# The Temple University Hospital Digital Pathology Corpus

*D. Houser[1], G. Shadhin[1], R. Anstotz[1], C. Campbell[1], I. Obeid[1], J. Picone[1]*
*T. Farkas[2], Y. Persidsky[2] and N. Jhala[2]*

1. The Neural Engineering Data Consortium, Temple University
2. Department of Pathology, Temple University
{devin.houser, tug69453, ryan.anstotz, christopher.campbell, iobeid, picone}@temple.edu
{tunde.farkas, yuri.persidsky, nirag.jhala}@tuhs.temple.edu

***Abstract*— Digital pathology is a relatively new field that stands to gain from modern big data and machine learning techniques. In the United States alone, millions of pathology slides are created and interpreted by a human expert each year, suggesting that there is ample data available to support machine learning research. However, the relevant corpora that currently exist contain only hundreds of images, not enough to develop sophisticated deep learning models. This lack of publicly accessible data also hinders the advancement of clinical science. Our digital pathology corpus is an effort to place a large amount of clinical pathology images collected at Temple University Hospital into the public domain to support the development of automatic interpretation technology. The goal of this ambitious project is to create a corpus of 1M images. We have already released 10,000 images from 600 clinical cases. In this paper, we describe the corpus under development and discuss some of the underlying technology that was developed to support this project.**

## I. Introduction

Pathology is a subset of medical science related to the cause, origin, and nature of disease [1]. A typical pathology laboratory workflow begins with a technician preparing the slide. This involves placing a tissue specimen on a glass slide, such as the specimen shown in Figure 1, and staining the slide for observation [2]. A board-certified pathologist views the stained specimen through an analog microscope to determine a diagnosis. The glass slides are then sorted by their respective tissue type and stored long-term often at an off-site location.


Figure 1. Example of a tissue biopsy specimen [2][3]

Digital pathology is the process of digitizing an analog image, so that images can be processed by a computer. Digitizing slides into whole slide images (WSI) provides both short-term and long-term benefits. Pathologists may provide real-time, remote analysis of the specimen and examine the sample with multiple pathologists simultaneously. Additionally, pathologists may retrieve existing digital WSIs through keyword searches and have access to the corresponding electronic medical records. Long-term advantages include the prevention of physical slide decay (i.e. stain discoloration, tissue degradation) over time [3].

WSIs are scanned using a digital slide scanner (DSS) and electronically stored. A typical scanned image is shown in Figure 2. With the advent of inexpensive digital storage, low-cost compute clusters, and cloud data storage, it is a cost-effective endeavor to maintain digital image archives of pathology slides. The ability to process slides digitally at a healthcare provider in real-time by using a DSS represents a transformative capability for clinical workflows.

Estimates indicate that approximately 10M pathology slides are observed each year in the United States. However, despite the existing volume of data, no comprehensive public WSI corpus exists. Currently, available resources such as the Cancer Genome Atlas (TCGA) Cancer Digital Slide Archive (CDSA) consists of WSIs on the scale of hundreds of slides per cancer type [4]. A study by Barker et al. [5] utilized the TCGA corpus and claimed machine performance that exceeded human performance. The dataset contained 604 WSIs of

Figure 2. Sample of a breast cancer case using H&E staining

two types of brain cancer: 364 glioblastoma multiforme and 240 lower grade glioma.

While access to public datasets is limited to hundreds of WSIs, private corpora, such as Philips and LabPON, contain WSIs on the scale of hundreds of thousands [6]. However, these private corpora are built and maintained to develop and deliver proprietary software to the end user. Access to any such proprietary corpus, if possible, often requires highly restrictive data licensing terms in addition to payment for distribution costs (which can be high in a clinical setting). Furthermore, distribution terms lead to long delays in acquiring and accessing the data.

Deep learning systems rely heavily on big data resources. Current public corpora, as previously outlined, are not adequate for supporting such technology development due to their small sample size. Therefore, the intent of this paper is to outline the implementation, impact, and goals of an NSF-funded Major Instrumentation Grant (NSF-MRI) [7] to develop a large, open source corpus of pathology images than can support state of the art machine learning research. Practicing clinicians will have unencumbered access to annotated WSI samples for educational purposes. Additionally, industry and research facilities will have big data resources for the development of deep learning algorithms and systems. We are collaborating with the Department of Pathology at Temple University Hospital (TUH) with a goal of scanning 1M slides from their vast archive of clinical data. This corpus, known as the TUH Digital Pathology Corpus (TUDP), will be released into the public domain over the next two years.

## II. Image Digitization

We use a Leica Biosystems Aperio AT2 high volume scanner [8], as shown in Figure 3, to scan our slides. Aperio's Scanscope Console software is used to scan images and eSlideManager (eSM) **Error! Reference source not found.** to database the images and manage metadata. This scanner is an industry-leading unit that includes a 40x scanning resolution and a z-stacking feature (up to 25 layers). A typical single specimen slide requires approximately 200 Mbytes of storage, though more complex images can increase in size to 1 Gbyte for multiple specimen slides (and up to 5 Gbytes for z-stacked images). The AT2 scanner can hold a total of 400 slides arranged in 40 carousels of 10 slides each. A period of approximately seven hours is required to scan 400 slides. This operation is run overnight, and the slides are organized using eSM the following day.

Figure 3. The Leica Aperio AT2 digital slide scanner

The scanning process is not as automated as one might expect. Before the scanner is set to conduct full scans overnight, pre-scan snapshots are taken of each slide. The duration of this process is approximately two hours per 400 slides. The snapshots are completed to allow for focus points to be placed on each snapshot and to specify the area of the slide that should be scanned. Fortunately, the software identifies focus points automatically, allowing the user to complete a quick review of the snapshot prior to processing.

In some cases, the software cannot locate enough focus points, and focus point data must be manually input. This event tends to occur with slides that are lightly stained, or slides that have a relatively high percentage of white space between tissue samples. Manual placement of focus points on the image is labor intensive; so, if many images require focus point placement, a considerable amount of labor is required. If the image does not contain sufficient focus points, the scanner will fail to process the image correctly and will not scan the slide. Of the 400 slides regularly set to scan overnight, approximately five slides, or about 2%, fail to scan. However, this number fluctuates depending on the quality of the slide stains. The slides that fail are reviewed and scanned again the following morning.

The Aperio AT2 scans the slide and creates a digital image stored in a proprietary file format known as a .svs file [10]. This file contains the raw image data and some metadata captured during the scanning process. A .svs file is a layered TIFF which uses JPEG 2000 compression to compress the actual image. The .svs file contains a thumbnail, which is a low-quality image of the slide, and the label image which is a low-resolution picture of the slide's label. Both are viewable when the slide is opened in Aperio ImageScope software [11]. Other information stored includes downsample and offset information.

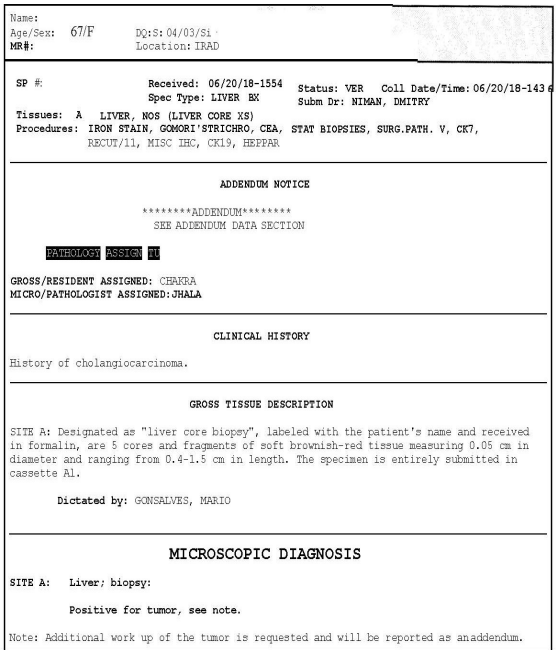
Fortunately, open source software tools such as [12] exist that allow these files to be viewed and manipulated. Though the WSIs could be converted to standard JPEG image files (.jpg), because the scanner scans at such a high resolution, the dimensions of a full-slide image exceed the limits of the JPEG format. Therefore, we are using .svs files as the primary filetype for the corpus because it is efficient and handles full resolution images.

## III. Data Anonymization

Once the scanner has finished digitizing the slides, the images are sorted into the corpus using the patient and sample number as descriptors. Under the Health Insurance Portability and Accountability Act (HIPAA) [13], patient identity must be kept entirely anonymous. To aid in this process, Protocol No. 24943 was approved by Temple's Institutional Review Board (IRB) [14]. This protocol ensures that all necessary measures are taken to protect research subject information. Because of this, the scanner resides at Temple University Hospital. The slides never physically leave the hospital for obvious reasons. The digital images are stored on a secure HIPAA network and remain on this network until the data is anonymized.

Patient data related to their identity must be removed through a process often referred to as deidentification or anonymization. A similar process to deidentify patient information was implemented by the Temple University Hospital EEG Seizure Corpus (TUSZ) [15]. TUH assigns patients a unique 8-digit Medical Record Number (MRN). A new randomized 8-digit key that maps to this MRN is created and used to anonymize the patient's identity. The mapping file information remains on the hospital's secure HIPAA network.

Each clinical case is accompanied by a report which provides information about the case. The first page of a typical anonymized report is shown in Figure 4. Information such as patient name, sex, age, MRN, and the date of the sample is located on the front page of the report. More specific details are found below, such as the clinical history of the patient, a gross description of the sampled tissue, and a medical diagnosis completed by a pathologist. TUH stores all the reports on their own database known as EPIC. Unfortunately, access to EPIC was restricted so each report is printed directly from the database for use in the TUH Digital Pathology Corpus and is scanned into PDF format. The reports are then converted to a Microsoft Word Document (.docx) and any patient data is removed. The anonymized patient report is then placed into the corpus that will be publicly released.



Name:
Age/Sex: 67/F DQ:S:04/03/Si
MR#: Location: IRAD

SP #: Received: 06/20/18-1554 Status: VER Coll Date/Time: 06/20/18-143 6
Spec Type: LIVER EX Subm Dr: NIMAN, DMITRY
Tissues: A LIVER, NOS (LIVER CORE XS)
Procedures: IRON STAIN, GOMORI'STRICHRO, CEA, STAT BIOPSIES, SURG.PATH. V, CK7, RECUT/11, MISC IHC, CK19, HEPPAR

ADDENDUM NOTICE

********ADDENDUM********
SEE ADDENDUM DATA SECTION

PATHOLOGY ASSIGN ME

GROSS/RESIDENT ASSIGNED: CHAKRA
MICRO/PATHOLOGIST ASSIGNED:JHALA

CLINICAL HISTORY

History of cholangiocarcinoma.

GROSS TISSUE DESCRIPTION

SITE A: Designated as "liver core biopsy", labeled with the patient's name and received in formalin, are 5 cores and fragments of soft brownish-red tissue measuring 0.05 cm in diameter and ranging from 0.4-1.5 cm in length. The specimen is entirely submitted in cassette A1.

Dictated by: GONSALVES, MARIO

MICROSCOPIC DIAGNOSIS

SITE A: Liver; biopsy:

Positive for tumor, see note.

Note: Additional work up of the tumor is requested and will be reported as anaddendum.

Figure 4. An excerpt from a typical anonymized report

The .svs files also contain a snapshot of the glass slide label. Figure 5 shows an open .svs file in the Aperio ImageScope software. The label usually contains the patient name as well as the sample number. To protect patient information, this layer has been removed in the public releases, leaving only the image itself in the file and various metadata created by the scanner. The metadata in the .svs files contains information that was recorded automatically, such as the scanner id and information about the image unrelated to the patient. This information can be obtained using Openslide.

## IV. Data Organization

Aperio's eSM comes with three options for organizing data, each one designed for a different purpose: Research, Educational, and Clinical. The Research option was chosen as the initial method of data organization due to its structure. The top level of the research category contains broad tissue type categories called Cases. These include Breast, Prostate, Gastrointestinal, Head/Neck, Pulmonary, and Endocrine. Inside each of these cases are more specific specimens. Each specimen is characterized by a specific clinical case. Specimens contain all the image files associated with the case as well as the clinical report in the form of a word document. A screenshot of the slide level organization is shown in Figure 6.

When the Aperio AT2 scanner is configured to scan images directly to the eSM database, all images are stored

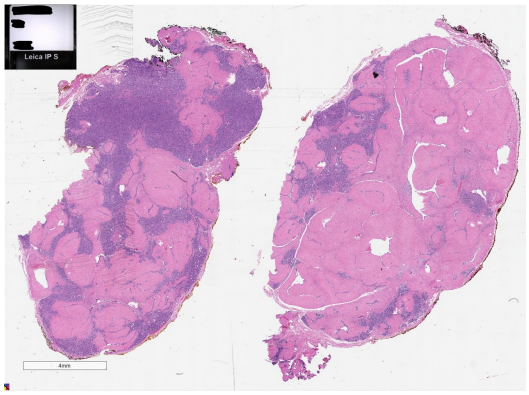



Figure 5. A typical scanned image before being anonymized

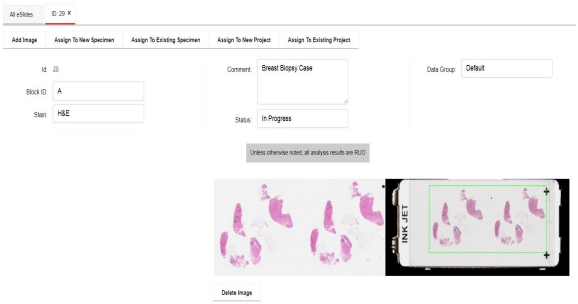



Figure 6. A screenshot of a databased image

in single directory on the HIPAA secured server. Each image is given a sequential name starting at one. This is far from a working naming convention for a public release, so a separate corpus was created for easy distribution. The top level of the corpus describes the broad tissue type of the cases underneath. The tissue types are the same as used in eSM. Under the tissue types, the cases are then sorted by the patients unique MRN. This additional level of organization allows for individual cases to be associated with specific patient information, which has proven to be valuable for age and sex-based studies.

Below the MRN level, the cases are separated by their respective case number, with the slide images and report in this case number directory. An image name is created by using the case number and other identifiers from the slide, such as the tissue site and code signifying the type of cut that was used. Samples are numbered chronologically starting with the year the sample was obtained, and then, the sample number. For instance, the first sample of 2002 is labeled s02_00001. These sample numbers will also be randomized in the public release to further protect the patient's private information.

Below is an example full filename for a typical image in the public corpus:

/breast/12345678/0s02_00001/0s02_00001_0a001_lvl0001_s00.svs

The first directory describes the type of tissue followed by the 8-digit patient MRN. The next fields are the clinical case name and the actual filename. The image name begins with a repeat of the case name, followed by the site code, 0a001, which is the specific area of the original biopsy. Following this is the type and number of the cut, lvl0001. This is done to ensure that each image has a unique filename.

The most common codes for the type of cut are level (lvl), deep (dep), and recut (rct). Level is a standard code for the site, given to slides to separate them from samples of the same tissue site. Deep slides are created if the initial sample is not sufficiently clear and a deeper tissue sample is taken. Recuts are completed only if the deeper tissue still has not given the required level of information prescribed. The last three characters of the image filename, s00, are used to separate images that have the same exact information, which is not common.

Most of the slides scanned from each case are stained with hematoxylin and eosin (H&E), which is the most common staining procedure currently used in pathology. A few immunochemically stained slides are included in the corpus, but many of these slides are not present in the cases we receive at TUH. There are plans in place to begin scanning more immunochemically stained slides in the future, but as of now they are difficult to catalog. The wide variety of staining types in these slides is difficult to characterize using the current naming convention.

## V. Computing Infrastructure

To support the long-term goals of this project and digitize the contents of all the digital pathology slides, it is necessary to implement a storage architecture that extends far beyond the demands of ordinary computing. It is a consequence of the extremely high resolution at which digital pathology images are scanned. Acquiring the raw storage to support the corpus becomes non-trivial. Cloud storage is enticing technology for many problem spaces but is not without its drawbacks. In addition to virtualization overhead, the cost of cloud storage at scale can be prohibitive. For example, Amazon's S3 storage pricing is $0.021/GB-month (using their U.S. East coast pricing) [16], which at 1.5 petabytes becomes $34,500/month. It is primarily for this reason that alternatives to cloud storage were considered for this purpose.

For this type of corpus, which could find uses in both research and clinical settings, storage must be highly available, so that operations performed on the data (e.g. research, annotations, diagnosis) do not suffer from high latency or low throughput, which would cause application performance to suffer, or even render an application utilizing the corpus unusable. Another requirement, especially in a clinical setting, is robustness with respect to physical hardware failure.

For diagnostic or clinical purposes, the data stored would be HIPAA-protected data, meaning that any disk or hardware failure resulting in data loss could be catastrophic. For this reason, the system needs to be able to withstand multiple distributed hardware failure events and maintain data integrity.

Respecting the constraints given by this application domain, we have developed the large fileserver architecture shown in Figure 7. Since the storage is distributed among a network of machines, it is both extensible and fault-tolerant, since each machine can be assigned an identical backup machine so that should an entire machine fail completely, data is recoverable from the backup. At the filesystem level, a ZFS filesystem [17] was used. ZFS uses its own RAID implementation that

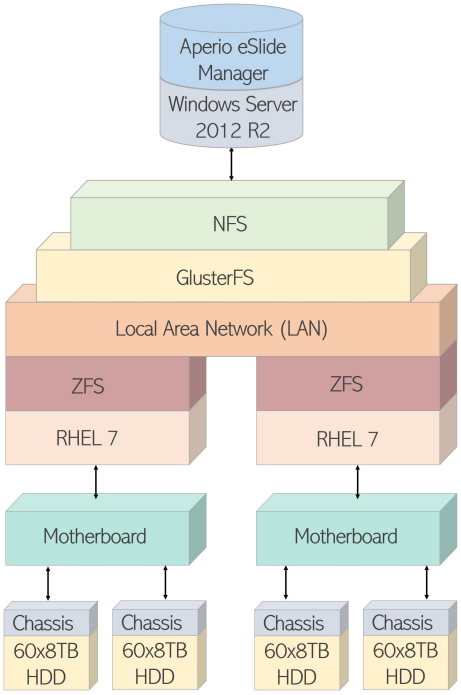



Figure 7. A large fileserver architecture

allows for a certain number of disks (depending on the configuration) to fail without losing data.

At the highest level of abstraction is the GlusterFS filesystem [18], which is a network-aware meta-filesystem capable of taking existing filesystems on multiple machines and combining them. It is at this level that machines could be configured to exist in mirrored pairs, so that any data written to one machine is automatically replicated on its mirror. The physical implementation of this system is shown in Figure 8. The two petabyte machines and the server hosting eSM are connected on a VLAN separate from our HPC cluster.

## VI. Annotation

Currently, the TUDP Corpus contains over 11,000 slides from 597 unique patients and 651 different cases. Most of the slides are composed of urology/prostate, breast and gastrointestinal cases. The remaining slides fall under smaller categories, such as pulmonary, head/neck and endocrine. These statistics are compiled in Figure 9. The slides included in the "miscellaneous" category do not belong to any of the broader categories due to a lack of information about the sample.

The integration of the corpus into the workflow at TUH relies on Aperio's eSM software **Error! Reference source not found.**. Cases are assigned to an individual pathologist for viewing. Pathologists can quickly analyze and annotate clinical slides that require review. eSM is a browser-based tool, so no additional software needs to be

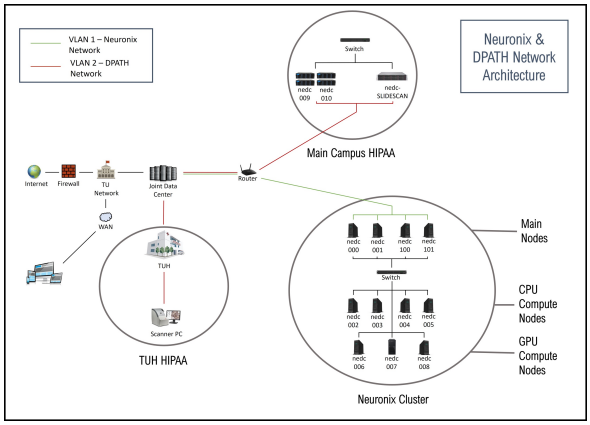



Figure 8. A HIPAA-compliant network architecture

installed on a client's computer – a major consideration when interfacing with hospital IT organizations. The tool is essentially a GUI front-end that connects to a back-end SQL database. This database contains manually entered patient and case information as well as the paths to the images, allowing for easy access and storage of the data on a large scale.

The .svs files can be opened and viewed using free software created by Aperio called ImageScope [11]. From within the ImageScope software, pathologists annotate regions of the image and add comments. After the annotations are complete, the annotation information, which includes coordinate information defining the location of the region that was annotated, shape information describing the boundaries of the region, and text containing the specific comments made by the pathologist.

The Aperio ImageScope software allows several shapes to be created to define the region of interest. The available shapes include a rectangle, ellipse and a freehand polygon form. Pathologists often argue that rectangular and elliptical shapes are not useful to signify regions of interest since diseased regions are usually irregular in shape. Because of this, the freehand annotations will be extremely important. Freehand annotations ensure that diseased tissue is fully encompassed and there is no room for misinterpretation of the affected region.

While 400 slides can be scanned every night by the Aperio AT2 scanner, a single pathologist would not be able to annotate that volume of slides in a day. According the code of federal regulations, the maximum number of slides that a can be viewed in an 8-hour work day is 100 slides (approximately 5 minutes per slide) **Error! Reference source not found.**. Therefore, the proposed 1M image database would require over 80,000 hours of annotation time, or 41 person-years. This is unacceptably high, so we are exploring ways to reduce this time, as we have done with other bioengineering applications. Not all slides require 5 minutes, however, and we are still early

| Spec Type | Number of Patients | Number of cases | Number of slides | Avg number of slides per Patient |
|---|---|---|---|---|
| Breast | 292 | 304 | 3,224 | 11.04 |
| Endocrine | 7 | 11 | 136 | 19.43 |
| Gastrointestinal | 109 | 109 | 1,966 | 18.04 |
| Head, Neck | 4 | 4 | 75 | 18.75 |
| Pulmonary | 4 | 4 | 55 | 13.75 |
| Urology | 180 | 194 | 5,083 | 28.24 |
| Miscelleneous | 8 | 25 | 684 | 85.50 |
| Total | 604 | 651 | 11,223 | 18.58 |

Figure 9. Some preliminary corpus statistics

in the annotation process, so these estimates are likely to change.

## VII. Summary

To share clinical cases with colleagues without WSI digitization, the glass slides must be physically transported to separate locations for review, which is highly time-consuming and not economical. The ability to scan slides into digital images and place them on a secure HIPPA network accessible to pathologists promotes remote accessibility and collaboration on patient cases. Pathology is a field where practitioners do routinely consult colleagues for second opinions, and WSI digitization is enhancing these collaborations. Educational practice also stands to gain from having access to a large, open source pathology corpus that can be easily searched using unstructured queries.

The Clinical Laboratory Improvement Amendments [20] require that histology slides be kept for a minimum of ten years after the date of examination. This requires hospitals to maintain massive physical archives of these slides. Slides are often stored off site and are not accessible without transportation back to the hospital, creating an additional layer of inefficiency and increasing operating costs. WSI digitization greatly enhances the value of these archives for clinical practice, research and teaching. The physical storage required to manage such a large archive electronically is now affordable and can be implemented using the relatively simple architecture described in this paper.

Our primary goal in this project, which began in October 2017, is to scan and annotate 1M slides in three years. Each image will have an associated annotation file and clinical report. The clinical reports only provide information for the whole case, with little to no information given on individual WSIs. Therefore, the associated annotation file is vital to understand the location(s) of diseased regions on the image. Similarly, the precise pixel data within the annotation file will aid in machine learning applications by allowing algorithms to associate specific areas with disease.

This corpus will enable the development of state-of-the-art machine learning systems, which require vast amounts of training data, for pathology applications such as cancer identification. This will, in turn, enable the development of software systems to assist diagnosis and accelerate the diagnostic process. This will ease pathologist workloads, which is important since projections are that there will be a shortage of pathologists in the coming decade **Error! Reference source not found.**.

Since this project is in its early stages, we welcome you to monitor our progress via the project web site, *www.isip.piconepress.com/projects/nsf_dpath* and provide feedback. The data and resources described in this paper will be available from this site.

## Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1726188. 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] H. Sattar, *Fundamentals of Pathology: Medical Course and Step 1 Review*, 8th ed. Chicago, Illinois, USA: Pathoma, LLC, 2017.

[2] G. Rolls, "An Introduction to Specimen Preparation," Leica Biosystems, 2018. *https://www.leicabiosystems.com/pathologyleaders/an-introduction-to-specimen-preparation/.*

[3] "Mounting Tissue Sections," National Diagnostics, 2011. *https://www.nationaldiagnostics.com/histology/article/mounting-tissue-sections.*

[4] D. Gutman, J. Cobb, D. Somanna, Y. Park, F. Wang, T. Kurc, J. Saltz, D. Brat, L. Cooper, and J. Kong, "Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data," *J. Am. Med. Informatics Assoc.*, vol. 20, no. 6, pp. 1091–1098, 2013.

[5] J. Barker, A. Hoogi, A. Depeursinge, and D. Rubin, "Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles," *Med. Image Anal.*, vol. 30, no. 1, pp. 60–71, 2016.

[6] H. Driessen, "Philips and LabPON plan to create world's largest pathology database of annotated tissue images for deep learning," Philips, 2017. *https://www.philips.com/a-w/about/news/archive/standard/news/press/2017/20170306-philips-and-labpon-plan-*

*to-create-worlds-largest-pathology-database-of-annotated-tissue-images-for-deep-learning.html*.

[7] 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, National Science Foundation, January 1, 2018.

[8] "Aperio AT2 – High Volume, Digital Whole Slide Scanning," Leica Biosystems, 2018. *https://www.leicabiosystems.com/digital-pathology/scan/aperio-at2/.*

[9] "Aperio eSlide Manager – Complete Digital Pathology Management Software," Leica Biosystems, 2018. Available: *https://www.leicabiosystems.com/digital-pathology/manage/aperio-eslide-manager/.*

[10] "Aperio Format," OpenSlide, 2018. *https://openslide.org/formats/aperio/.*

[11] "Aperio ImageScope - Pathology Slide Viewing Software," Leica Biosystems, 2018. *https://www.leicabiosystems.com/digital-pathology/manage/aperio-imagescope/*.

[12] M. Satyanarayanan, A. Goode, B. Gilbert, J. Harkes, and D. Jukic, "OpenSlide: A vendor-neutral software foundation for digital pathology," *J. Pathol. Inform.*, vol. 4, no. 1, p. 27, 2013.

[13] R. Brzezinski, *HIPAA Privacy and Security Compliance - Simplified: Practical Guide for Healthcare Providers and Managers 2016 Edition*, 3rd ed. Seattle, Washington, USA: CreateSpace Independent Publishing Platform, 2016.

[14] "Development of Pathology Data Corpus." Institutional Review Board (IRB) - Protocol 24943, Temple University, 2018.

[15] I. Obeid and J. Picone, "The Temple University Hospital EEG Data Corpus," *Front. Neurosci. Sect. Neural Technol.*, vol. 10, p. 00196, 2016.

[16] "Simple Monthly Calculator," Amazon Web Services, 2018. *http://calculator.s3.amazonaws.com/index.html*.

[17] J. Bonwick, M. Ahrens, V. Henson, M. Maybee, and M. Shellenbaum, "The Zettabyte File System," in *Proceedings of the 2nd Usenix Conference on File and Storage Technologies*, 2003, pp. 1–13.

[18] B. Depardon, G. Le Mahec, and C. Seguin, "Analysis of Six Distributed File Systems," Institut National de Recherche en Informatique et en Automatique (INRIA), Lyon, France, 2013. *https://hal.inria.fr/hal-00789086/document*

[19] 42 CFR 493.1274 - Standard: Cytology. United State of America: Cornell Law School, 2018.

[20] "Clinical Laboratory Improvement Amendments," 1988. *https://wwwn.cdc.gov/clia/Regulatory/default.aspx.*

[21] N. Jhala, "Digital pathology: Advancing frontiers," in *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2017.