There are three significant outcomes from this project:

First, an Aperio AT2 scanner has been deployed at Temple Hospital for four years and is being used on a daily basis to scan slides for our corpora. It is also used by pathologists to support their tumor board reviews and personal research projects. We scan approximately 400 slides per day and also support the daily operational needs of the pathologists. We have integrated the scanner into our computer cluster - neuronix.nedcdata.org. Neuronix is unique in that we have successfully connected three highly secure networks (Temple’s main campus public network and two HIPAA-secured private networks) so that we can easily move data from hospital operations to research, once the data is properly deidentified. The networks are interconnected via VPNs and satisfy all necessary security protocols. All students and staff allowed access to these networks have undergone thorough HIPAA and CITI training and are certified to handle this type of data. The deidentified data is released via an anonymous rsync server and directly browsable from our organization’s web site. Our ability to access Temple Hospital data directly, deidentify it under an approved IRB protocol, and release it as open source data makes this facility fairly unique.

Second, we have scanned over 80,000 slides including a 13,000+ slide subset provided by Fox Chase Cancer Center (FCCC, affiliated with Temple). We have released an annotated breast tissue corpus containing 3,505 slides. A rough breakdown of pathologies included in the corpus include: Breast Tumor (6%), Urinary Prostate (36%), Gastrointestinal (18%), Lymph Nodes (8%) and Other (32%). The FCCC subset contains 18 types of tissue (38.5% prostate, 16.5% gynecological, 45% other). We will release the remaining unannotated data as well in Spring 2022 to support machine learning experiments in unsupervised and self-supervised learning. Annotations are released in two formats: csv and xml. The former is the preference of researchers; the latter is what is generated from our annotation tools.

Corresponding patient reports and supporting immunohistochemical stains are provided as part of the metadata available with the corpus. The microscopic diagnoses given by the primary pathologist in these reports detail the pathological findings within each tissue site, but not within each specific slide. The microscopic diagnoses informed our annotation process. Further differentiation of cancerous and precancerous labels, as well as the location of their focus on a slide, was accomplished with supplemental immunohistochemically (IHC) stained slides. When distinguishing whether a focus is a nonneoplastic feature versus a cancerous growth, pathologists employ antigen targeting stains to the tissue in question to confirm the diagnosis. For example, a nonneoplastic feature of usual ductal hyperplasia will display diffuse staining for cytokeratin 5 (CK5) and no diffuse staining for estrogen receptor (ER), while a cancerous growth of ductal carcinoma in situ will have negative or focally positive staining for CK5 and diffuse staining for ER. Many tissue samples contain cancerous and non-cancerous features with morphological overlaps that cause variability between annotators. The informative fields IHC slides provide could play an integral role in machine model pathology diagnostics.

Third, we developed an automated image classification system based on deep learning principles that classifies regions with an accuracy of at least 75% for all but one label (nneo). This system is available as open source software. An overview of the system is provided in the attached image labeled Figure 4. The core deep learning technology used is based on Convolutional Neural Networks (CNNs) using a configuration known as ResNet18. Transfer learning is used to augment the training process. Traditional image processing methods such as binarization, dilation, erosion, hole filling and region removal based on ranks are used to segment the images, rank areas and form regions of interest, which are known as patches. This not only improves performance but also greatly reduces the overall processing time. A postprocessing step is applied in which patch detections are compared with their neighbors and adjusted using a majority voting scheme of its 8 neighbors to make the final decision. This approach removes a minority of mis-detected patches and greatly increases accuracy.

A summary of performance is shown in the attached image labeled Figure 5. This provides performance in the form of a confusion matrix. Whole slide classification performance (e.g., cancer vs. non-cancerous) is very high – 94% accuracy. The code is written in Python and makes use of the Pytorch machine learning package.