## What is the impact on the development of the principal discipline(s) of the project?

There are three types of impact this project has had on the principal discipline(s), which we consider to be machine learning and pathology. First, through the procurement of a digital slide scanner, the first such device available at Temple University Hospital, we were able to introduce the Pathology Department to the benefits of digital scanning. They now routinely use digital images of slides for their regular reviews, such as their weekly tumor board reviews. We have supported scanning of the materials presented at these reviews and captured these as part of our corpus. Beyond the clear impact this has on patient care, it represents a chance for clinical care at these institutions to take a step forward.

Second, through the development of our open source corpora, we have made available several landmark corpora. Consistent with our overall philosophy of open source resources at the Neural Engineering Data Consortium, we have made our corpora available in an unencumbered manner. Gaining access to the data is immediate and instantaneous (though we collect user information so we can track our subscribers and keep them informed of recent developments). In an era where HIPAA concerns and intellectual property issue still prevent many researchers from releasing data, our data is open source and available. We have released two important subsets of data: (1) the Breast Tissue Subset, which consists of 3,505 slides, and (2) The Fox Chase Cancer Center Biorepository data, which is close to 14,000 slides (and still growing). The Breast Tissue subset contains annotations. Typically, five to ten regions on each slide have been annotated for various pathologies. Further, the data was split into three subsets: training, development, and evaluation data. This facilitates machine learning research and allows researchers to directly compare results in a scientific fashion. Deidentified pathology reports are also available for each session represented in the corpus, which makes it a valuable corpus for natural language processing researchers.

The Fox Chase data represents an extensive repository curated by Fox Chase Cancer Center over the past decade. Though the data is not annotated, each slide does contain a summary diagnosis, which allows whole-slide machine learning experiments, often associated with end-to-end deep learning system development, to be conducted. (We are in the process of annotating the breast tissue subset of this data as part of a follow-on project.)

We are also in the process of releasing about 100,000 slides that have been scanned but not annotated. This is taking much longer than expected, but the data should become available by Summer 2022. We are committed to making this release. We hope, with follow on funding in place, we can continue the development of annotated versions of this data.

The third type of impact this project has made is the release of a software system that automatically localizes patches of images determined to be cancerous. This Python-based system is highly accurate on both slide-level and patch-level classification. It is currently specific to breast tissue but can be easily retrained for other classification tasks. The scripts are easy to follow and self-documenting. We also provide online support for customers via our email support listserv [help@nedcdata.org](mailto:help@nedcdata.org). The code is well organized and straightforward. This makes it very easy for new researchers, such as students, to enter the field since they have access to data and an end-to-end system that shows them how to set up and run experiments on the data.

The overarching goal for this MRI project was to create momentum in the field for digital pathology and automated analysis tools based on machine learning. Not only have we had an immediate and direct impact on Temple University and Fox Chase Cancer Center, but as a result of this project we are building large-scale collaborations between medical schools, machine learning researchers, natural language processing researchers and clinicians. Cross-fertilization of knowledge will undoubtedly create new research opportunities.

## What is the impact on other disciplines?

Due to the high resolution nature of digital pathology images, this type of data is of great value to the image processing community at large. It is not trivial to apply traditional deep learning approaches on this type of data since the data does not easily fit into computer memory. Data flow, algorithm design and computational hardware all become important issues when attempting to process this kind of data. Hence, establishing this task as a community-wide challenge task, which we are working on, will be an important driver for the machine learning community in general.

Similarly, the presence of large amounts of annotated data makes the data extremely useful for educational purposes. We expect faculty will integrate this data in their teaching tools at medical schools and clinical practices.

Further, the existence of high-quality annotations can serve as a catalyst to unify and normalize annotation conventions and techniques for manual interpretation. We have seen this first-hand in this project as we have had pathologists from several institutions give us feedback on our annotations and experienced variations in conventions across institutions.

## What is the impact on the development of human resources?

Our pool of undergraduate workers come from a diverse set of fields including bioengineering, biochemistry, electrical and computer engineering, neuroscience, and mechanical engineering. These students don’t always get exposed to computer-based technologies such as machine learning. The students involved in this project acquire a wide range of skills including Linux computing, Python programming and data science in general. Often this makes them want to pursue careers in the computational sciences upon graduation. It definitely impacts their view of medicine and science even if they stay in their discipline.

We currently have three graduates of this project in medical school. We also have had several undergraduates decide not to pursue medical degrees in favor of data science positions. The engineers often end up being highly sought after in fields like embedded systems that require a mixture of hardware, software, and data science skills.

On top of these technical skills, all our students participate in the organization and execution of an annual conference we host – the IEEE Signal Processing in Medicine and Biology Symposium (IEEE SPMB). This provides them an opportunity to publish, present and participate in a professional conference. We work closely with our students on their technical communications skills, which helps them significantly upon graduation.

## What was the impact on teaching and educational experiences?

The data and resources we develop are used in a wide range of classes including Engineering Computation I (ECE 1111), Engineering Computation IV (ECE 4822), Senior Design (ENGR 4196/4296) and Introduction to Machine Learning and Pattern Recognition (ECE 8527). Students often use our data to test data-intensive programs or conduct machine learning experiments on data sets we prepare for educational purposes.

All of our project-related material are distributed via our web site – *isip.piconepress.com*. This web site is one of the longest-running open source web sites in engineering. It is well known for distributing a wide range of educational materials in addition to data and software. For example, the ISIP Machine Learning Demo (IMLD), available at *www.isip.piconepress.com/projects/imld*, is one such example. This is a tool that allows students to visualize and experience machine learning technology.

## What is the impact on physical resources that form infrastructure?

As part of this project, we have constructed the first computer network at Temple University that spans the main campus, a HIPAA-restricted research network, and the hospital's operational HIPPA-controlled network. It took a long time to make this happen, but this is one of the few networks of its types. Our students can move data across each of these networks without being physically located at the hospital. This is not a great technical achievement, but simply a demonstration of persistence. It involved solving a number of complex firewall issues. It puts us in a very unique position with respect to data collection. More about this network can be found at: *www.isip.piconepress.com/projects/neuronix*.

## What is the impact on institutional resources that form infrastructure?

As mentioned before, the ability for our students to access hospital data directly via a HIPAA-secured VPN was a first at Temple University. We spent a significant amount of time negotiating this with Temple University and Fox Chase Cancer Center administration. The net result is that we have created a proof of concept arrangement from which many researchers at Temple can benefit from. We continue to consult and advise other research projects on these protocols and capabilities.

## What is the impact on information resources that form infrastructure?

As explained above, we have worked very closely with campus and hospital IT teams to achieve a very unique virtual network that allows us to move HIPPA-compliant data. Many other research groups at Temple will benefit from the path we have opened with this project.

## What is the impact on technology transfer?

We have released a working implementation of a classification system that can be used to automatically interpret medical images. The software is flexible enough to be applied to problems beyond digital pathology. For example, we are currently involved in a project where we are adapting the software to analysis of 3D computational biology images. We expect researchers will be very interested in learning about the pragmatic approaches we have taken to process large numbers of high resolution images.

## What is the impact on society beyond science and technology?

As part of our outreach activities, we are engaging high school students in year-long mentoring activities and summer internships within our lab. For these students, many of whom come from impoverished area in the city of Philadelphia and under-resourced high schools, it is a valuable opportunity to gain exposure to the expectations of a college education, and how to go about gaining admittance into a good university.

## What percentage of the award's budget was spent in a foreign country?

None.