Preview of Award 1726188 - Final Project Report

Cover |
Accomplishments |
Products |
Participants/Organizations |
Impacts |
Changes/Problems
| Special Requirements

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	1726188
Project Title:	MRI: High Performance Digital Pathology Using Big Data and Machine Learning
PD/PI Name:	Joseph Picone, Principal Investigator Tunde Farkas, Co-Principal Investigator Iyad Obeid, Co-Principal Investigator Yuri Persidsky, Co-Principal Investigator
Recipient Organization:	Temple University
Project/Grant Period:	01/01/2018 - 12/31/2021
Reporting Period:	01/01/2021 - 12/31/2021
Submitting Official (if other than PD\PI):	N/A
Submission Date:	N/A
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	N/A

Accomplishments

*** What are the major goals of the project?**

As shown in Figure 001 (attached), there are three major goals of this project:

- Phase 1: Hardware Acquisition
- Phase II: Data Development
- Phase III: Algorithm Development

This first phase consists of hardware procurement, development and installation. This required managing a complex multi-organizational relationship between the vendor (Aperio/Leica Biosystems), the Temple Hospital Information Technology group (TUHS-IT) and Temple University Main Campus Information Technology (TUMC-IT). Subgoals for this phase of the project included:

- Procurement of Scanner
- Installation and Verification
- User Training
- Procurement of Network Storage
- Installation and Verification
- Final Hardware Certification

The second phase of the project consisted of data development. This involved working closely with the vendor to certify and integrate the slide scanner hardware and software. Subgoals for this phase of the project included:

- Preliminary Archival Scanning
- User Acceptance Testing
- Production Archival Scanning
- Workflow Integration
- User Acceptance Testing
- Production Scanning
- IRB Application
- Preliminary Database Release

- User Acceptance Testing
- Production Database Release

The third phase of the project consisted of algorithm development. This involved the development of a deep learning-based system to automatically classify data. Subgoals for this phase of the project included:

- Pilot Experiments
- System Tuning
- System Performance Analysis
- Physician Feedback
- Final System Performance Evaluation
- Physician Acceptance Testing

The releases of this data and software should create significant community-wide interest in digitial pathology. This is a field that is beginning to take off as a number of commercial vendors have released products during the duration of the this project. It is our hope that the data developed in this project can serve several purposes, including creating industry-accepted benchmarks for performance. To that end, another important goal of the project was to create a community-wide presence for this project and the resources developed.

*** What was accomplished under these goals and objectives (you must provide information for at least one of the 4 categories below)?**

Major Activities:

The first two goals of the project collectively result in our ability to collect data. Over the course of the project we have digitized over 100,00 slides. We have not released all of this data yet because we are still in the process of renaming and organizing it. However, as described in various sections of this report, we have released two important subsets.

The average image size is 372 Mbytes, which confirms the fact that these are extremely high-resolution images. The total size of our released data is close to 7 Terabytes. We have reconfigured our web server to support the dissemination of these large datasets. While this activity is a detail beyond the scope of the report, it was an important step to allow people to efficiently download such a large amount of data.

We have also trained undergraduate students to annotate breast tissue data. We have believe we have this process under control and are generating high quality annotations. We have since expanded into the next largest subset of the data - urinary prostate slides. To do this, we had to learn how to classify these slides. We created a collaboration with the Pathology Department at Fox Chase Cancer Center (FCCC), a center of excellence for cancer treatment in the U.S.

The third task, the development of software to automatically classify slides, has resulted in a release of an open source system that we refer to as a Digital Pathology Cancer Recognition System (DPATH-CRS). An artificial neural network (ANN) is used to classify the regions of the digital pathology slides. An ANN tries to assimilate the recognition ability of a pathologist. ANNs have many free parameters that need to be tuned to the specific applications - a process is called training. Then this trained ANN is evaluated by classifying some unseen data. If it can guess the correct classes, then it can be employed as a diagnostic system, and if not, the training process should be repeated.

We designed a five-step ANN training and evaluation pipeline:

1. PATCH EXTRACTION

A digital pathology slide file contains several images with different resolutions but from a single specimen. Annotators, that are trained to diagnose the regions, attach the labels to some selected regions, but not all parts of a slide. Pathologists usually look at the very subtle details of the slide to diagnose the region. Therefore, the processing is done on the highest resolution which is called level 0. Level 0 contains a very large image, such as 50Kx50K pixels.

Patches are square-shaped images that are cropped from some parts of level 0 and processed as an atom. This means that a patch is assumed to contain a single class, not different classes, and the diagnosis system assigns a single label to it. Annotated areas are made with freehand drawing tools, so they can have any shape. Also, all patches in an annotated area have a single label. A rolling window is used and the window size is selected to be 256x256, with a step size of 256 in both directions. Every window that has more than 50% overlap with the annotated area is assumed to be an acceptable patch and is stored as a 256x256 TIFF image in the corresponding patches directory.

At this point, all annotated areas from all slides are broken down into patches, and patches are stored in the corresponding directories.

2. PATCH DISTRIBUTION

Patch distribution is necessary to allocate different slides to the training, development, and evaluation steps of an ANN. Two considerations make the intelligent distribution of the patches crucial. First, the ratio of patches in every class in comparison to the total patches must be kept approximately similar. It means that if the ratio of patches of class 1 in the training dataset is 10%, then it is expected to have the same percentage in the development and evaluation dataset. Second, to have an open-loop (blind) development and evaluation, there must be different slides in different datasets and no similar slides should be in two different datasets.

To satisfy these two constraints, a greedy algorithm is used. Since solving this problem in an optimal way is an NP-hard problem, to make the whole process fast, we have to relax some constraints. The second constraint is crucial, so we relaxed the first one. Instead of maintaining all classes' ratios fixed, just the classes with the smallest percentages are tried to be kept fixed. Hence, a greedy search algorithm tries to maintain the percentage of classes in order. It distributes the slides that contain the rarest class with predefined ratios. Then, it distributes the slides with the second rare classes, again with predefined ratios, but with priority given to training, development, and evaluation, respectively.

3. TRAINING

At this point, the slides that are used for training are known, but the number of patches in the classes varies, resulting in an unbalanced dataset. Two approaches to solve this problem were examined: weighted loss and random sampling. The weighted loss function assumes that the prior probabilities of classes in the datasets are proportional to the number of patches in that class. This proportional weighting can be used in either original format or after a softmax operation.

The second approach is based on random sampling. The idea comes from the observation that most of the patches in the outnumbered classes are very similar, especially in the patterns. So, it is rational that instead of using all the available patches in these outnumbered classes, a limited number of them are selected randomly and then train the ANN on this random subset of that class, while keeping all the rare classes as they were. This approach is called random sampling.

Combining the random subset approach with the weighted loss function solves the highly unbalanced dataset issue. But how effective is random sampling. To find out, the ANN is trained on some random subset of the training dataset, then the one that has the best loss on the development set is selected as the best trained ANN. Since the random subset of the training dataset is usually much smaller, the training finishes fast, but the development process takes more time because the ANN is examined on all the patches on the development dataset.

An ANN model needs to be predetermined before the training. The convolutional layers in the image analysis with the deep learning methods are proven to be very effective for making a pyramidal representation of an image. Therefore, most image analysis systems use some convolutional filtering in a hierarchical order or in a parallel format for extracting the patterns in different resolutions and making a new set of features. Then with the help of one or two fully connected layers, they classify these features. We follow the same approach, but use a pre-trained ANN. This approach is known as transfer learning. We use a very well-known, simple, and very effective kind of these network which is called Residual Network or ResNet.

4. DECODING

Given an unannotated whole slide, it is cropped into too many patches. These patches are shown to the neural network and the ANN classifies every patch, which is called decoding. In the decoding phase, some minor errors can be pruned. We can define a minimum size for every region. For example, a cancerous region in the size of a single patch should not be surrounded by all the normal neighbors and it is expected that some similar patches make a large area. So, a 2D moving average filter or a 2D Gaussian filter with a predefined radius is used to estimate the center patch and this estimation is being compared with the decoded label. If they are close to each other, nothing is changed, but if they are too different, the estimated output of the filter is used instead of the decoded decision.

5. EVALUATION

Evaluation needs annotated data. So, only the decoded parts that contain annotations are used. A confusion matrix is generated that is very essential evaluating the effectiveness of this system.

Specific Objectives:

The specific objectives for the project were essentially: (1) procure and deploy research instrumentation that supported the digitization of pathology slides; (2) the development of a large corpus of annotated pathology slides that will support machine learning; and (3) develop deep learning software that automatically classifies and segments pathology slides, so that pathologists can integrate these classifications into their workflows, thereby improving their efficiency.

Significant Results:

The major findings for this project ae summarized in the poster presentation attached as Figure 2. We have accomplished each of the specific objectsives described above:

(1) Research Instrumentation: we have deployed a digital slide scanning system at the hospital that provides the ability to scan, archive and retrospectively search large numbers of images. The scanner is capabable of sanning 400 slides in approximately 8 hours. The file store can hold one petabyte of data (approximately 2.7M images). The computers can be accessed remotely using VPN technology. This allows us to scan and remotely access all the tools and data while maintaining HIPAA compliance.

(2) Corpus Development: We have released two major subsets of the data, and plan to release almost 100,000 slides in total. The Breast Tissue subset is now being actively used in research. The FCCC Corpus is also beginning to be used to baseline clinical systems.

(3) Deep Learning Software: We have released a deep learning classification system that is based on a ResNet18 architecture. It is capable of classifying patches at accuracies above 80%, and can categorize whole slides at an accuracy of approximately 94%.

Key outcomes or Other achievements:

Pathologists at Temple Hospital are now beginning to examine our automatically generated annotations of slides using the retrospective search capability of our tools. They are in the early stages of determining the clinical usefulness of the system. Initial feedback is promising. Follow-on funding is allowing us to do a more detailed analysis of the clinical usefulness of these tools.

*** What opportunities for training and professional development has the project provided?**

Students involved in this project often enter the project will little or no software experience and no background in patholgy. We have employed students from a wide range of backgrounds including biochemistry, bioengineering, electrical and computer engineering, computer science and data science. The impact of this project on these students cannot be understated since we teach them virtually everything they need to know about topics ranging from the basic science of pathology to software and computing.

We provide extensive training on Linux-based computing, software development in C++ and Python, and a variety of project management tools. We also encourage our students to publish at our annual conference, write book chapters, etc. We work closely with them training them how to write technical publications.

The skills they develop allow them to transition into summer internships in industry, where they gain even more valuable skills relevant to their careers. We spend a lot of time assisting these students in their job searches. Many of our students continue into internships involving clinical experiences, continue into medical school, move into research positions at start ups or transition into postdocs and staff positions at local healthcare institutions.

Their combination of computing skills and domain knowledge make them highly sought after. We cannot understand the impact this kind of funding has had on their careers.

*** Have the results been disseminated to communities of interest? If so, please provide details.**

In addition to distributing information through a listserv that we use to communicate with subscribers of our resources, we maintain a project web site and disseminate data,resources, and software through our digital pathology web site: https://isip.piconepress.com/projects/nsf_dpath/. Users can gain access to the data by using a completely automated sign up process located at:

https://isip.piconepress.com/projects/nsf_dpath/html/request_access.php

The Downloads tab on this web site leads visitors to the resources we have developed. There are currently two corpora: the Temple University Hospital Breast Tissue Corpus and the Fox Chase Cancer Center Corpus. We have also released our baseline deep learning automated annotation system that is implemented in Python.

We also host an annual Institute of Electrical and Electronics Engineers (IEEE) conference at which we publish our related research. There are have been contributions related to this project during the duration of the project. Both our graduate students and undergraduates participate in this conference by presenting papers and posters on our EEG research. The archives of the conference are indexed in IEEEXplore. The conference web site is located here: https://www.ieeespmb.org/ and is freely available. Since the pandemic began, we have been conducting the conference as a virtual conference and publishing videos of the talks in addition to the papers. Conference participation has doubled when we moved to a virtual format. Interest in pathology remains strong and we use the conference to educate our subscribers about our resources.

Supporting Files

Filename	Description	Uploaded By	Uploaded On
figure_001.pdf	The original timeline for the project. The project was extended through 2021 on a no-cost extension due to delays encountered due to COVID.	Joseph Picone	05/23/2022
figure_002.pdf	An overview of the significant results obtained in the project.	Joseph Picone	05/24/2022

Products

Books

Book Chapters

Shawki, N., Shadhin, M. G. M., Elseify, T., Jakielaszek, L., Farkas, T., Persidsky, Y., Jhala, N., Obeid, I., & Picone, J. (2020). The Temple University Digital Pathology Corpus. *Signal Processing in Medicine and Biology: Emerging Trends in Research and Applications 1.* 1. Obeid, I., Selesnick, I., and Picone, J.. Springer. New York City, New York, USA. 67. Status = PUBLISHED; Acknowledgement of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1007/978-3-030-36844-9.

Inventions

Journals or Juried Conference Papers

View all journal publications currently available in the NSF Public Access Repository for this award.

The results in the NSF Public Access Repository will include a comprehensive listing of all journal publications recorded to date that are associated with this award.

Wevodau, Z. and Doshna, B. and Jhala, N. and Akhtar, I. and Obeid, I. and Picone, J.. (2021). The Temple University Digital Pathology Corpus: The Breast Tissue Subset. *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. 1 (1) . Status = Deposited in NSF-PAR doi:https://doi.org/10.1109/SPMB52430.2021.9672275 ; Federal Government's License = Acknowledged. (Completed by Picone, Joseph on 04/23/2022) Full text Citation details

Campbell, C. and Mecca, N. and Duong, T. and Obeid, I. and Picone, J.. (2018). Expanding an HPC Cluster to Support the Computational Demands of Digital Pathology. *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. 1 (1) 01 to 03. Status = Deposited in NSF-PAR doi:https://doi.org/10.1109/SPMB.2018.8615614 ; Federal Government's License = Acknowledged. (Completed by Picone, Joseph on 01/26/2021) Full text Citation details

Hunt, I. and Husain, S. and Simon, J. and Obeid, I. and Picone, J.. (2019). Recent Advances in the Temple University Digital Pathology Corpus. *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. 1 (1) 1 to 4. Status = Deposited in NSF-PAR doi:https://doi.org/10.1109/SPMB47826.2019.9037859 ; Federal Government's License = Acknowledged. (Completed by Picone, Joseph on 01/26/2021) Full text Citation details

Houser, D. and Shadhin, G. and Anstotz, R. and Campbell, C. and Obeid, I. and Picone, J. and Farkas, T. and Persidsky, Y. and Jhala, N.. (2018). The Temple University Hospital Digital Pathology Corpus. *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*. 1 to 7. Status = Deposited in NSF-PAR doi:10.1109/SPMB.2018.8615619 ; Federal Government's License = Acknowledged. (Completed by Picone, Joseph on 11/03/2019) Full text Citation details

Licenses

Other Conference Presentations / Papers

Other Products

Other Publications

Patent Applications

Technologies or Techniques

Thesis/Dissertations

Websites or Other Internet Sites

Supporting Files

Filename	Description	Uploaded By	Uploaded On
web_site.pdf	A screenshot of the main page of the project web site that is used to disseminate information.	Joseph Picone	05/24/2022

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Picone, Joseph	PD/PI	4
Farkas, Tunde	Co PD/PI	1
Obeid, Iyad	Co PD/PI	1
Persidsky, Yuri	Co PD/PI	1
Khalkhali, Vahid	Graduate Student (research assistant)	1
Shawki, Nabila	Graduate Student (research assistant)	1
Alexander, Carmel	Undergraduate Student	0
Battalora, Leo	Undergraduate Student	2
Brown, Angela	Undergraduate Student	1
Campbell, Chris	Undergraduate Student	2
Cap, Thao	Undergraduate Student	2
Cason, Noah	Undergraduate Student	1
Chhin, Sidney	Undergraduate Student	1
Doshna, Ben	Undergraduate Student	2
Duong, Thuc	Undergraduate Student	1
Elseify, Tarek	Undergraduate Student	2
Gorecki, John	Undergraduate Student	1
Gutierrez, Fernanda	Undergraduate Student	1
Houser, Devin	Undergraduate Student	2
Hunt, Isabel	Undergraduate Student	2
Husain, Saiyeda	Undergraduate Student	2
Jakielaszek, Luke	Undergraduate Student	2
Jean-Paul, Shmyrde	Undergraduate Student	2
Juarez Mendez, Javier	Undergraduate Student	2
Lebo, Savannah	Undergraduate Student	1

Name	Most Senior Project Role	Nearest Person Month Worked
Liang, Dennis	Undergraduate Student	2
Lloyd, Landen	Undergraduate Student	1
Makholia, Paras	Undergraduate Student	1
Mecca, Nicholas	Undergraduate Student	2
Meng, Phuykong	Undergraduate Student	1
Mir Muzamil, Temory	Undergraduate Student	1
Paroya, Abraham	Undergraduate Student	1
Shadhin, Golam	Undergraduate Student	1
Shaw, Skyler	Undergraduate Student	2
Simons, Julien	Undergraduate Student	2
Teperov, Josh	Undergraduate Student	2
Tulin, Nikita	Undergraduate Student	1
Vorwick, Lynn	Undergraduate Student	1
Wevodau, Zoe	Undergraduate Student	2
Xu, Emily	Undergraduate Student	1

Full details of individuals who have worked on the project:

Joseph Picone
Email: joseph.picone@gmail.com
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 4

Contribution to the Project: Technical lead and project manager. Oversaw all aspects of the project including administration, execution, publication and dissemination.

Funding Support: None.

Change in active other support: No

International Collaboration: No
International Travel: No

Tunde Farkas
Email: tunde.farkas@temple.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Domain expertise.

Funding Support: None.

Change in active other support: No

International Collaboration: No
International Travel: No

Iyad Obeid
Email: iobeid@temple.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Management of Institutional Review Board issues. Bioengineering subject matter expertise. DevOps support and software dissemination.

Funding Support: None.

Change in active other support: No

International Collaboration: No
International Travel: No

Yuri Persidsky
Email: yuri.persidsky@temple.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: TUHS/LKSOM Pathologist. Subject matter expertise. Administration of Temple Hospital issues such as computer accounts, lab space, IT, etc.

Funding Support: None.

Change in active other support: No

International Collaboration: No
International Travel: No

Vahid Khalkhali
Email: vahid.khalkhali@temple.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 1

Contribution to the Project: algorithm software; baseline annotation system development.

Funding Support: None

International Collaboration: No
International Travel: No

Nabila Shawki
Email: nabila.shawki.68@gmail.com
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 1

Contribution to the Project: algorithm software; baseline annotation system development.

Funding Support: None

International Collaboration: No
International Travel: No

Carmel Alexander
Email: carmel.alexander@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 0

Contribution to the Project: Data annotation

Funding Support: None.

International Collaboration: No
International Travel: No

Leo Battalora
Email: leo.battalora@gmail.com
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: system administration (IT); hardware installation and deployment; IT support for personnel working on the project.

Funding Support: None.

International Collaboration: No
International Travel: No

Angela Brown
Email: angela.brown0003@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Data annotator and database developer.

Funding Support: None.

International Collaboration: No
International Travel: No

Chris Campbell
Email: tuf46810@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: system administration (IT); hardware installation and deployment; IT support for personnel working on the project.

Funding Support: None.

International Collaboration: No
International Travel: No

Thao Cap
Email: thao.cap@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: Data annotator and database developer.

Funding Support: None.

International Collaboration: No
International Travel: No

Noah Cason
Email: noah.cason@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Data annotator and database developer.

Funding Support: None.

International Collaboration: No
International Travel: No

Sidney Chhin
Email: sidney.chhin@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Data annotator and database developer.

Funding Support: None.

International Collaboration: No
International Travel: No

Ben Doshna
Email: benjamin@doshna.net
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: Data annotator and database developer.

Funding Support: None.

International Collaboration: No
International Travel: No

Thuc Duong
Email: tug98850@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: system administration (IT); hardware installation and deployment; IT support for personnel working on the project.

Funding Support: None.

International Collaboration: No
International Travel: No

Tarek Elseify
Email: tug35668@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: web design; baseline system development

Funding Support: None.

International Collaboration: No
International Travel: No

John Gorecki
Email: john.gorecki@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Software developer and System Administrator

Funding Support: None.

International Collaboration: No
International Travel: No

Fernanda Gutierrez
Email: fernanda.gutierrez@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Data annotator and database developer.

Funding Support: None.

International Collaboration: No
International Travel: No

Devin Houser
Email: tuf89323@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: Managed data collection at the hospital; operate the digital scanning equipment for hospital pathologists; organize medical reports; anonymize the data.

Funding Support: None.

International Collaboration: No
International Travel: No

Isabel Hunt
Email: tuj85563@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: Managed data collection at the hospital; operate the digital scanning equipment for hospital pathologists; organize medical reports; anonymize the data.

Funding Support: None other than this award.

International Collaboration: No
International Travel: No

Saiyeda Husain
Email: tuh11491@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: Managed data collection at the hospital; operate the digital scanning equipment for hospital pathologists; organize medical reports; anonymize the data.

Funding Support: None

International Collaboration: No
International Travel: No

Luke Jakielaszek
Email: tug52339@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: web design; baseline system development

Funding Support: None

International Collaboration: No
International Travel: No

Shmyrde Jean-Paul
Email: tuh26880@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: computer system administrator / IT support

Funding Support: None

International Collaboration: No
International Travel: No

Javier Juarez Mendez
Email: javier.juarez.mendez@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: Software developer and System Administrator

Funding Support: None.

International Collaboration: No
International Travel: No

Savannah Lebo
Email: savannah.lebo@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Data annotator and database developer.

Funding Support: None.

International Collaboration: No
International Travel: No

Dennis Liang
Email: tul14986@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: Web developer responsible for the dissemination of information about the project

Funding Support: None

International Collaboration: No
International Travel: No

Landen Lloyd
Email: tun33314@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Data annotator and database developer.

Funding Support: None.

International Collaboration: No
International Travel: No

Paras Makholia
Email: tuj81485@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Web developer responsible for the dissemination of project information

Funding Support: None

International Collaboration: No
International Travel: No

Nicholas Mecca
Email: tuf89560@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: system administration (IT); hardware installation and deployment; IT support for personnel working on the project.

Funding Support: None

International Collaboration: No
International Travel: No

Phuykong Meng
Email: phuykong.meng@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Software Developer

Funding Support: None.

International Collaboration: No
International Travel: No

Temory Mir Muzamil
Email: mirmuzamil_temory@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Data annotator and database developer.

Funding Support: None.

International Collaboration: No
International Travel: No

Abraham Paroya
Email: paroyaabraham@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Software developer and System Administrator

Funding Support: None.

International Collaboration: No
International Travel: No

Golam Shadhin
Email: tug69453@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Managed data collection at the hospital; operate the digital scanning equipment for hospital pathologists; organize medical reports; anonymize the data.

Funding Support: None

International Collaboration: No
International Travel: No

Skyler Shaw
Email: tuk32485@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: system administration (IT); hardware installation and deployment; IT support for personnel working on the project.

Funding Support: None.

International Collaboration: No
International Travel: No

Julien Simons
Email: tuk39975@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: Managed data collection at the hospital; operate the digital scanning equipment for hospital pathologists; organize medical reports; anonymize the data.

Funding Support: None.

International Collaboration: No
International Travel: No

Josh Teperov
Email: josh.teperov@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: system administration (IT); hardware installation and deployment; IT support for personnel working on the project.

Funding Support: None.

International Collaboration: No
International Travel: No

Nikita Tulin
Email: tug47034@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Database developer, data annotation, operations manager

Funding Support: None

International Collaboration: No
International Travel: No

Lynn Vorwick
Email: tug70217@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: software engineer; image recognition system development

Funding Support: None

International Collaboration: No
International Travel: No

Zoe Wevodau
Email: tuk59951@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: Database developer / data annotator

Funding Support: None

International Collaboration: No
International Travel: No

Emily Xu
Email: tuk90002@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Data annotator and database developer.

Funding Support: None.

International Collaboration: No
International Travel: No

What other organizations have been involved as partners?

Name	Type of Partner Organization	Location
Fox Chase Cancer Center (FCCC)	Other Nonprofits	Philadelphia, PA

Full details of organizations that have been involved as partners:

Fox Chase Cancer Center (FCCC)

Organization Type: Other Nonprofits
Organization Location: Philadelphia, PA

Partner's Contribution to the Project:
In-Kind Support
Collaborative Research

More Detail on Partner and Contribution: FCCC, as described in the report, helped train our students on data annotation and will be providing data (pathology slides).

Were other collaborators or contacts involved? If so, please provide details.
Nothing to report

Impacts

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. 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?

No funds were expended in a foreign country.

Changes/Problems

Changes in approach and reason for change
Nothing to report.

Actual or Anticipated problems or delays and actions or plans to resolve them
Nothing to report.

Changes that have a significant impact on expenditures
Nothing to report.

Significant changes in use or care of human subjects
Nothing to report.

Significant changes in use or care of vertebrate animals
Nothing to report.

Significant changes in use or care of biohazards
Nothing to report.

Change in primary performance site location
Nothing to report.

Special Requirements

Responses to any special reporting requirements specified in the award terms and conditions, as well as any award specific reporting requirements.