My Desktop
Prepare & Submit Proposals
Prepare Proposals in FastLane
New! Prepare Proposals (Limited proposal types)
Proposal Status
Awards & Reporting
Notifications & Requests
Project Reports
Submit Images/Videos
Award Functions
Manage Financials
Program Income Reporting
Grantee Cash Management Section Contacts
Administration
Lookup NSF ID

Preview of Award 1726188 - Annual 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 - 12/31/2020
Reporting Period:	01/01/2018 - 12/31/2018
Submitting Official (if other than PD\PI):	Joseph Picone Principal Investigator
Submission Date:	01/03/2019
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	Joseph Picone

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

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

Major Activities:

Phase I – Hardware Acquisition: The major challenge in the first year was to install the computer hardware. There were two major components to this: the Aperio/Leica slide scanner and associated software, and the four computer systems required to support this hardware. Before we purchased the Aperio scanner, we did an extensive product search. There were two major vendors of this technology: Aperio, who was recently purchased by Leica Biosystems, and Philips. After about 6 months of evaluations with both vendors we selected Aperio because of their industry-leading position, their open file formats and their large capacity (400 slides). An order for this equipment was placed at the end of 2017, and the equipment arrived in early January 2018.

In parallel, we designed and purchased a network of computers to support the technology. This hardware consisted of two 1-petabyte file servers, which we planned to install on the Temple University Main Campus HIPAA network for security reasons, an app server that functions as a web server and delivers the ImageScope application, and a PC that hosts the scanner. The app server and scanner PC reside on the hospital HIPPA network, which is used by the hospital for its daily work. Hence, great care was taken not to interfere with normal hospital operations. A network diagram is shown in Figure 002 (attached).

Though we collected requirements from Aperio prior to purchasing this equipment and cleared the network design with all IT groups at Temple that had a stake in this project, getting these computers to talk to one another turned out to be an enormous effort. This was the first time such a thing had been done at Temple University, so there were a

large number of people (over 20) directly involved in this project. After 9 months of work, we were able to reach a solution that was acceptable to all parties, and the systems were placed into operation.

We implemented virtual file systems on top of the basic file systems provided by these systems so that we can easily backup and manage the disks. We use a combination of the open source tools Gluster and ZFS to allow us to have a virtual RAIDed file system that spans all disks and allows one machine to mirror the other, so backups are instantaneous and automatic.

A major reason for placing the app server on the TUHS HIPAA network was so that pathologists can use the same sign-on that they use for their regular work. This single sign-on makes it very easy for them to access the systems. Unfortunately, it took quite a bit of behind-the-scenes work to make this happen.We made all of this work without compromising the strict security procedures that are associated with these secure networks.

Phase II – Data Development: Once the hardware was stabilized, we needed to train our undergraduate student workers on operation of the scanner. This began with a short training session that was conducted during the equipment installation in January 2018. We learned how to physically operate the scanner (e.g., loading slides into the slide trays), how to enter and organize the images into the database tool, and most importantly, how to annotate images. This allowed us to proceed with digitizing slides and learning the nuances of the Aperio software.

One June 27, 2018, Aperio visited Temple to provide official training. They presented a one-day course that covered all aspects of their technology. This included a one-hour session for all the Temple pathologists to introduce them to the process of viewing and annotating images. Since the hardware was now relatively stable, we felt it was time to discuss integrating the technology into their workflow.

Over the last 6 months of 2018, we began ramping up the digitization process. By the end of 2018, we had digitized over 15,000 images from over 700 cases. We have been focusing on three types of cases: breast cancer, gastrointestinal and urinary. We are using breast tumor cases as a pilot corpus to benchmark our automatic classification software based on advice given by our co-PIs.

We have also been digitizing slides to support tumor boards that the pathologists conduct weekly. This is the first step in demonstrating the value of digital imaging to the pathologists. We prepare a selected set of images for them, and they use the ImageScope software to display and manipulate the images during their board meeting. We have been supporting this activity for several months now and it is a significant milestone because it represents the first insertion of the technology into their workflows.

There are two important observations about the process that we are following. First, part of the scanning process requires previewing the slides and setting “focus points.” Our technician typically loads the scanner with 400 slides, previews them to make sure they will scan properly, and then run the scanner overnight. It takes about 6 hours to scan 400 slides, so it makes sense to let this run overnight. Focus points are a critical part of how the scanner adjusts its focus to optimally scan a particular section of a sample. Aperio uses an automated algorithm to suggest focus points, but these often have to be adjusted manually if the slides are old (e.g., slides that are older than three years and candidates for purging), noisy (e.g., discoloration or stray marks) or complex (e.g., a breast tumor slide that has multiple images on a single slide).

We find that it is not uncommon to have 10% of the slides in a batch of 400 slides not scan properly. Fortunately, the Aperio software makes it very easy to determine which slides did not scan properly, and this does not terminate the process. However, this

means we have to manually re-scan these slides the following morning, which reduces the throughput we can achieve.

Second, the quality of the slides of course influences the reliability of the scanning process. Older slides tend to cause more problems. We try to scan slides before they are purged from the hospital archives so that we don't lose that data. These slides tend to cause more problems with scanning, so throughout is not as high as we would like due to the need to adjust the scanner for these slides. We have accumulated enough experience with the scanner that we are now better able to predict which slides will scan without problems.

Phase III – Algorithm Development: We also began developing a baseline classification system. This system is based on our state-of-the-art seizure prediction technology. We have trained a hierarchical deep learning system that uses a combination of Convolutional Neural Networks (CNNs) and Long Short-term Memory Networks (LSTM). We have optimized this system so that it does not load the entire image, or the entire set of images, into memory. This is an important issue because there is not enough memory on our GPU processors to hold these images. We have trained this system on synthetic data with known parameters and demonstrated that it properly classifies the images when the signal to noise ratio (SNR) is reasonable. We feel at this point the mechanics of the system is sound. It is written in the Python programming language and uses the open source toolkits Keras and TensorFlow.

Part of this development has included the development of a scoring process that properly calculates the degree of match between a reference event and a hypothesis when they don't overlap in space perfectly. This is a topic covered in great detail in our open source scoring software for EEG signals. EEG signals are one-dimensional signals – amplitude versus time. We have adapted this software to score images and modified the code that assesses overlaps between hypotheses to operate in two dimensions instead of one dimension.

We are now constructing a more realistic pilot data set that includes a single clearly visible artifact. We will use this data set to verify that the deep learning system is able to operate effectively on real image data.

Specific Objectives:

The specific objectives for the first year of the project are shown in Figure 001 (attached). There are four objectives that have been emphasized in this first year:

- Procure the instrumentation, which in this case consisted of computer hardware, install it, and resolve all associated networking issues.
- Master the Aperio ImageScope software installation, operation and training, including training of TUHS pathologists on its use.
- Develop a workflow that would allow us to scan a high volume of slides each week.
- Implement a Python-based deep learning infrastructure based on the open source packages known as Keras and TensorFlow that is capable of processing large amounts of images efficiently.

We have made significant progress on all of these objectives. Hardware deployment is ahead of schedule while the other objectives are on schedule.

Significant Results:

Significant outcomes from our first-year progress include:

- **Hardware Infrastructure:** We met our milestone of delivering 1 petabyte of storage for less than $50K. This is an unprecedented price/performance ratio for off-the-shelf storage solutions and is an important enabler of this project. Completing installation of this massive file store in a way that is acceptable to TUHS-IT is a major step forward for the project, since TUHS-IT is very careful about what computer systems are allowed to interact with the hospital's production computer network.
- **Software Infrastructure:** Installation and operation of the Aperio slide scanner was fairly smooth. However, getting the file servers, which reside on the TU Main

Campus HIPAA network, and the machines hosting the Aperio software and hardware, which reside on the TU Hospital HIPAA network, proved to be quite a challenge. Though we went through an extensive requirement gathering process prior to the purchase and cleared the requirements with both TUHS-IT and TUMC-IT, Aperio failed to disclose several software requirements of their systems. This included a requirement that the PC scanner talk to the app server via a service account. level of Computer hardware is installed and operational. The Aperio software requires a service account to be established between the two machines. This is an enormous security risk and required both machines to reside on the same network – the TUHS HIPAA network in this case. It took months to resolve this issue, but we eventually arrived at a solution that allows the file servers to be on the Main Campus HIPAA network.

- **Slide Scanning:** The scanner has a 400-slide capacity. It was our understanding that a slide could be scanned without human intervention. However, we discovered that the process for scanning a slide includes a preview process where the scanner automatically sets focus points. These focus points are critical to achieving a high-quality scan. Some percentage of the time, depending on the quality of the slides, the scanner fails to automatically set the focus points, and they must be manually adjusted. This significantly slows our ability to do bulk scanning because a typical batch of 400 slides results in approximately 10% of the slides failing to scan properly. We have to manually set focus points for these and rescan them, which is very time-consuming. Slides that contain multiple images, such as a breast biopsy, must also use focus points that are manually set.
- **Imaging Science:** The scanner has the ability to generate 5K x 5K images. This is pretty impressive. Unfortunately, this resolution exceeds the limits of a jpeg file (to our surprise). Hence, we are using Aperio's ScanScope Virtual Slide (.svs) format. Fortunately, this is an open standard and many open source tools exist that can read these images, including some Python libraries. A .svs file is essentially a layered or tiered TIFF image. Images range in size from about 50M to 500M depending on their complexity and the amount of compression that can be achieved. Images can be as large as 5G if z-stacking is used (the ability to construct a 3D scan of the slide but scanning it in multiple slices). However, this feature is not being used by our pathologists. If a typical image is 100M, our 1 Petabyte file server can hold 10M images. Hence, we should be able to store 1M images on our fileserver.
- **Application Development:** We have been optimizing our Python-based image analysis software to read images efficiently yet minimize the amount of physical memory used (storing these large images entirely in memory is not ideal). We have evaluated several approaches and have arrived at a solution that reads images a block of lines at a time. This alleviates I/O bottlenecks and yet maintains simplicity of the algorithm software. The inefficiency of Python is mitigated with this approach and we have no problem running multiple jobs on the cluster that process large number of images.

Key outcomes or Other achievements:

Key outcomes for the project include:

1. **Hardware Deployment:** the entire hardware suite, which includes two 1 Petabyte file servers, a web server that hosts the Aperio application, and a PC that hosts the scanner, has been purchased, installed and made operational. This required working through many extremely complicated IT issues involving network security.
2. **Synthetic Data:** we have created a small corpus of synthetic images that can be used to debug preliminary implementations of deep learning algorithms. Successfully processing synthetic data is often the first step in the development of machine learning algorithms.
3. **Pilot Data:** we have released pilot data corpus, referred to as v0.0.0, that contains 6 patients and 3 tumor types (breast, gastrointestinal, urology). We have solicited feedback from the community on the quality and format of the data.
4. **Baseline Classification System:** we have completed the development of the infrastructure for a baseline classification system that has been optimized to

efficiently process large volumes of images. We are in the process of overlaying state of the art deep learning algorithms on this system and evaluating them on a simple artifact classification task.

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

All PIs and senior staff on the project have been trained on the use of digital imaging in pathology, including learning how to use the Aperio image viewing software. Several pathologists are now using the system to present cases at their weekly tumor boards.

The project budget included support for undergraduate students. We have assigned three undergraduate students to this project. Working closely with TUHS pathologists, they have been trained on various aspects of digital pathology including operation of the scanner, annotation of images using the Aperio software and interpretation of the images for specific diseases such as breast tumors. Our Information Technology Team, which consists of four student system administrators and one student web designer, has also been trained on the basics of the project. In addition, several research students have been involved in the development of our baseline classification system.

We also hosted a training session for TUHS pathologists on July 27, 2018 conducted by Aperio/Leica. Users were trained on how to manage, view and annotate images using the Aperio software. This was well-attended by the Department of Pathology at Temple University and initiated the process of pathologists using the system in their daily work. For many, this was the first time they were using digital imaging.

*** How have the results been disseminated to communities of interest?**

Our database products are well-known and disseminated through our web site. We have created a project web site, *www.isip.piconepress.com/projects/nsf_dpath*, that is used to keep our users up to date on the project. A screenshot of the main page of the web site is show in Figure 003 (attached). We regularly post updates to this web site to inform users of the availability of new resources.

We also maintain a Google Groups listserv, nedc_tuh_dpath @googlegroups.com, that we use to disseminate project announcements. We are following a similar model to what we have used to inform people about our EEG resources. Our EEG listserv, *nedc_tuh_eeg@googlegroups.com*, currently has over 1,500 users. We expect our digital pathology listserv, to easily exceed this number of users once the database is in a regular release cycle, and once we have had several significant publications on the project.

We also host an annual conference, the IEEE Signal Processing in Medicine and Biology Symposium (IEEE SPMB), that we use to disseminate information about the project. At IEEE SPMB 2018, we gave an oral presentation on the project (see the publication listed in this report). We also presented a demonstration of the image analysis system and a pilot version of the corpus. Proceedings of IEEE SPMB are indexed in IEEE Xplore, one of the world's largest repositories of technical information in engineering and computer science. We are also working with the publisher Springer to publish a book containing expanded versions of selected full papers presented at the conference. Our paper on this project will be featured as one of the chapters in this book.

*** What do you plan to do during the next reporting period to accomplish the goals?**

There are three major goals for the second year of the project:

1. **Data Development:** continue scanning new pathology slides, reaching a goal of 500K images.
2. **Data Annotation:** create a consensus on annotation standards for the community and disseminate at least 100K annotated images.
3. **Algorithm Development:** establish performance benchmarks for our baseline classification system.

We are well-positioned to reach these goals in the second year of the project. Data acquisition and development is now a relatively smooth process. In the last quarter of the first year of the project, we have begun working closely with the pathologists on annotation standards and have documented our current approach in a publication described below. We have also completed the development of a baseline system and are now in the process of evaluating it on a small pilot corpus of annotated images.

Refer to Figure 001 (attached) for a more detailed view of the timeline for the project.

Supporting Files

Filename	Description	Uploaded By	Uploaded On
figure_001.pdf	Figure 001: Detailed Timeline	Joseph Picone	01/02/2019
figure_002.pdf	Figure 002: The Digital Pathology Computing Cluster	Joseph Picone	01/02/2019
figure_003.pdf	Figure 003: The Project Web Site	Joseph Picone	01/02/2019

Products

Books

Book Chapters

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.

D. Houser, M. M. Golmohammadi, R. Anstotz, C. Campbell, I. Obeid, J. Picone, T. Farkas, Y. Persidsky, and N. Jhala, “The Temple University Hospital Digital Pathology Corpus,” in Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium, 2018, pp. 1–7.. Status = AWAITING_PUBLICATION.

Licenses

Other Conference Presentations / Papers

Other Products

Other Publications

Patents

Technologies or Techniques

Thesis/Dissertations

Websites
Supporting Files

Filename	Description	Uploaded By	Uploaded On
07_spmb_l01_06.pdf	IEEE SPMB Paper	Joseph Picone	01/03/2019

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Picone, Joseph	PD/PI	2
Farkas, Tunde	Co PD/PI	1

Name	Most Senior Project Role	Nearest Person Month Worked
Obeid, Iyad	Co PD/PI	1
Persidsky, Yuri	Co PD/PI	1
Campbell, Chris	Undergraduate Student	1
Duong, Thuc	Undergraduate Student	1
Elseify, Tarek	Undergraduate Student	1
Houser, Devin	Undergraduate Student	1
Jakielaszek, Luke	Undergraduate Student	1
Mecca, Nicholas	Undergraduate Student	1
Shadhin, Golam	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: 2

Contribution to the Project: project management, computing infrastructure

Funding Support: None

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: data review and annotation

Funding Support: None

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: subject matter expertise, publication support, data management

Funding Support: None

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: data review and annotation, subject matter expertise

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: 1

Contribution to the Project: system administration (IT); hardware installation and deployment

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 support

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: 1

Contribution to the Project: web design; baseline system development

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: 1

Contribution to the Project: Managed data collection at the hospital

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: 1

Contribution to the Project: baseline system development (technology)

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: 1

Contribution to the Project: system administration, IT support

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: Supported data collection at the hospital

Funding Support: None

International Collaboration: No
International Travel: No

What other organizations have been involved as partners?
Nothing to report.

What other collaborators or contacts have been involved?
Nothing to report

Impacts

What is the impact on the development of the principal discipline(s) of the project?
Nothing to report.

What is the impact on other disciplines?

The pilot data we have released will be useful towards generating interest in this field.

What is the impact on the development of human resources?

A significant number of undergraduates have worked on this project, introducing them to STEM opportunities in bioengingeering and machine learning. Two of the students, who are pursuing degrees in computer science, have commented on how this project has positively impacted their education, both from a programming point of view and a basic science point of view.

What is the impact on physical resources that form infrastructure?

Creating the computer network described in this report set all sorts of precedents at Temple University. Hopefully, in the future, it will be easier for researchers to interact with Temple Hospital and gain access to valuable data.

What is the impact on institutional resources that form infrastructure?

Coordinating between three major IT organizations on campus certainly broke new ground in our ability to collaborate as a university.

What is the impact on information resources that form infrastructure?

Creating an environment that involves communications between three secure neworks broke a lot of new ground at Temple University.

What is the impact on technology transfer?

The use of .svs images rather than .jpeg images certainly forced us to rethink our software designs.

What is the impact on society beyond science and technology?
Nothing to report.

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.

Special Requirements

Responses to any special reporting requirements specified in the award terms and conditions, as well as any award specific reporting requirements.
Nothing to report.