Real-time Accurate Imaging Diagnosis

Zahi Fayad, Patricia Kovatch, Laurie Margolies, David Mendelson, David Yankelevitz
Icahn School of Medicine at Mount Sinai

Response to NIH RFI on Science Drivers Requiring Capable Exascale HPC

The Specific Scientific Challenge

Healthcare providers need to determine the right test at the right frequency for optimal early detection of breast cancer, with limited false positives. Today, breast radiologists synthesize data from imaging informatics, the radiology subspecialty that includes Picture Archiving and Communication System (PACS), electronic medical records, structured reporting, computer assisted diagnosis, natural language processing (NLP), archiving, radiation dosimetry, data mining products, peer review, the exchange of health information and real-time education, to make the best diagnosis possible. As the breast imaging profession positions itself for the next twenty years of innovation in the era of big data, structured reporting and the ability to create and populate large databases that allow for data mining are critical to improve diagnosis and treatment. The ability for computer-aided diagnosis software to more quickly analyze and refine data from a much larger dataset would be transformative for patient care. For instance, if we could identify tumors with a high level of accuracy in real-time, it would revolutionize healthcare.

Developing highly accurate machine learning algorithms to perform feature extraction will require exabyte sized image repositories and exascale computing for data analysis. Correlations of specific image features would be matched with similar images from others along with genetic and other medical record information. We use all modalities, even for a simple study, asking questions such, "Does ultrasound or tomosynthesis add benefit to high risk breast screening if an MRI was done? Who needs a mammogram and at what frequency? Who needs supplemental ultrasound or MRI? More tests? Fewer tests?"

High dimensional data analysis could help answer these questions tailored to each patient. By performing some of this analysis in advance, we can work towards a real-time diagnosis suggested by software. To ensure that this computer-aided diagnosis software is highly accurate, we would need image, genetic and medical record data from tens to hundreds of millions of patients, thus requiring exabyte sized data warehouses and analysis engines. Ultimately, we want to mine previously uncollectable data to determine new, true risk factors beyond the traditional ones such as how much you walk from your cell phone, radiation exposure from flying, second hand or primary smoke exposure, food, medication, genes beyond BRCA1 and 2 and pTen and others. This would enable us to perform meaningful data mining and to improve accuracy and specificity need to amalgamate images from many institutions along with genetic and lifestyle and other risk factor data. When have all this data combined then it can be mined for the ultimate goal of preventing cancer or employing much more effective screening algorithms.

The Potential Impact

More accurate, real-time identification and diagnosis of lesions will enable faster treatment for afflicted patients and reduce patient stress for healthy patients. Coupled with genetic, medical record and other data from the digital universe, doctors will be empowered to provide more precise treatments for specific patients. The reduction of unnecessary biopsies and related tests will save significant amounts of healthcare dollars for patients as well as for local, state and federal governments.

The Specific Limitations of Existing HPC

This capability is currently limited by (1) the accuracy and complexity of the algorithms learning to automatically identify tumors, (2) the scale of the available images the algorithms need to learn from, and (3) the lack of correlation with other high dimensional data sets such as medical records and

genomics data. Existing centers do not have the exabyte storage capacity to host the images or the computing power to perform the necessary exascale data analytics to correlate them genetic and medical record data. Scalable software stacks and sophisticated machine learning algorithms for imaging are also missing.

Computational Parameters Expected in 2025

At ISMMS, we already store over 10 million images in all modalities (x-ray, ultrasound, MRI, mammography, CT Scan, MRI, PET, etc.) in over 4 petabytes of storage for over seven million patients since 2003. As imaging modalities become more complex and accurate, the file sizes for each modality will grow. We estimate that an average patient has 20 images taken over their lifetime representing 1 TB/patient. If we wanted to correlate images with genomic and medical record data from 1 million patients through President Obama's Precision Medicine Initiative, then we would need an exabyte of storage to hold all of this data.

Other Capabilities Needed

To achieve our stated goal of improved diagnosis, we will need access to an order of magnitude more images in order to train algorithms to successfully and accurately identify tumors. To do this, there will need to be a national repository for images, genetic information and medical records and/or a mechanism for sharing them in a decentralized way so that all researchers will be able access the data for additional studies. We need software that facilitates communication and integration of patient data from numerous, geographically distributed sources. The repository and software could be linked to President Obama's million person precision medicine initiative for additional scientific insight and analysis.