White Paper

AI in Healthcare

Smart Infrastructure Choices Increase Success

Shilpi Ahuja and Mike McNamara, NetApp
Brad Genereaux, NVIDIA
November 2019 | WP-7314

In partnership with

Abstract

From improving patient outcomes to expanding the reach of medical expertise to reducing costs, the promise of artificial intelligence in healthcare is immense. However, significant challenges stand in the way of successful AI deployment, in both the research lab and the clinic.

This white paper examines three important areas in which AI will play a significant role: medical imaging, digital pathology, and genomics. It explores challenges in each of these fields, discusses the critical role that data plays, and describes possible approaches to address computing and data storage requirements.

TABLE OF CONTENTS

1 The Promise of AI for Healthcare 3

1.1 AI Challenges in Healthcare 3

1.2 What This White Paper Covers 4

2 Medical Imaging 4

2.1 AI in Medical Imaging 5

2.2 Data Management in Medical Imaging 6

2.3 Research and Clinical Challenges in Medical Imaging 7

3 Digital Pathology 7

3.1 AI in Computational Pathology 8

3.2 Data Management in Digital Pathology 8

3.3 Research and Clinical Challenges in Digital Pathology 9

4 Genomics 9

4.1 AI in Genomics 9

4.2 Data Management in Genomics 10

4.3 Research and Clinical Challenges in Genomics 11

5 NetApp and NVIDIA Solutions for AI in Healthcare 12

Where to Find Additional Information 12

LIST OF TABLES

Table 1) Comparison of AI in medical imaging, digital pathology, and genomics. 12

LIST OF FIGURES

Figure 1) AI encompasses machine learning and deep learning and is closely associated with big data. 3

Figure 2) Pipeline stages for deep learning. 6

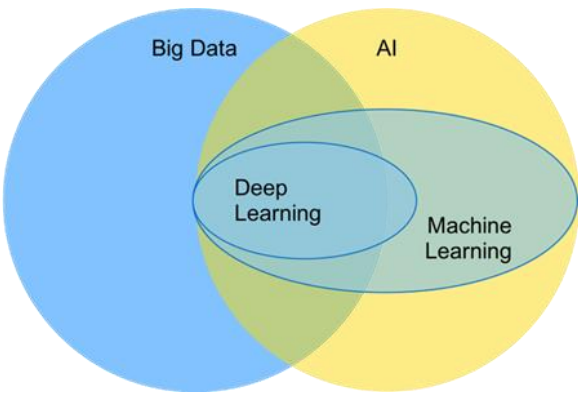
1 The Promise of AI for Healthcare

The promise of artificial intelligence (AI) is greater in healthcare than in almost any other industry. From improving patient outcomes to expanding the reach of medical expertise to reducing costs, the potential benefits are immense. AI is already helping to improve the accuracy of disease diagnosis and opening the door to precision medicine—the identification of the most effective treatment for a patient based on genetic, environmental, and other factors—with benefits that include earlier disease detection, tailored treatment plans, reduced radiation doses, and much more.

However, AI efforts in healthcare to date have barely scratched the surface of what will eventually be possible. The healthcare industry remains significantly behind other industries in AI adoption. This lag is in large part due to data privacy and access issues.

When we talk about AI in healthcare, we typically mean the application of deep learning algorithms using neural networks as well as other machine learning algorithms to healthcare data. (See Figure 1.) These algorithms require large amounts of high-quality data during training to achieve high accuracy, making data a fundamental consideration for AI in healthcare—as it is in all industries. Healthcare institutions face significant challenges in applying AI to the exabytes of data that they generate, most of which remains locked in and difficult to extract because of privacy issues.

Figure 1) AI encompasses machine learning and deep learning and is closely associated with big data.



1.1 AI Challenges in Healthcare

A healthcare AI project faces many of the same challenges as projects in other industries, including:

- Identifying the problem and the best way to attack it
- Collecting and managing massive amounts of data
- Finding and hiring people with the necessary data science and AI expertise
- Identifying the necessary compute and other resources for the AI project
- Transitioning a trained algorithm from proof of concept to production
- Training staff to take advantage of AI results, and integrating results into existing or new workflows

However, healthcare faces some additional challenges:

- **Data privacy.** Maintaining patient privacy and adhering to regulatory requirements (such as HIPAA in the United States and GDPR in Europe) is a paramount concern that can affect both how data is managed and how algorithms are trained. Patient confidentiality must be maintained while keeping track of data provenance for regulatory compliance. These requirements make it difficult to share raw data, which must be de-identified before it can be used for AI training. However, it's not possible to simply delete all identifying information in all cases. For example, to apply AI to tumor growth, it's necessary to associate all the studies from each individual patient.

- **Data specificity.** Healthcare data across different institutions—or even across different instrumentation—may not be strictly comparable. For example, an algorithm trained to identify tumors from MRI scans based on a particular dataset may not be as accurate when deployed in a different institution where the equipment is different or is calibrated differently.
- **Budget limitations.** Healthcare IT operations often face significant budgetary and other constraints. When it comes to capital outlays, IT has to compete for funds with other priorities. Targeted use of cloud resources can help with budget limitations, assuming that data privacy concerns can be addressed. On-premises investments in infrastructure for AI should be targeted to deliver the greatest possible leverage.
- **Research versus clinical needs.** Researchers and clinicians have different needs for AI. Ideally, researchers focus on training accurate models that can then be applied in a clinical setting to improve patient care, but there are gray areas that can create significant overlap.

Beyond these technical challenges, human safety is of utmost importance for healthcare organizations. It can be difficult to discern how AI algorithms reach their decisions, and a mistake can literally mean the difference between life and death, so many clinicians are justifiably skeptical.

Organizations should expect to expend significant time and effort to establish trust in AI technology, meet data integration challenges, and address privacy and security concerns.

1.2 What This White Paper Covers

This paper focuses on three prominent use cases of AI in healthcare:

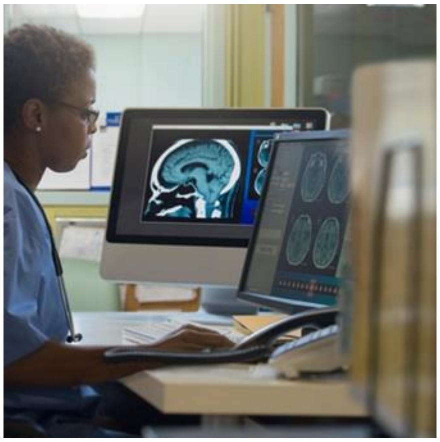
- Medical imaging
- Digital pathology
- Genomics

The use of AI in these three areas has not only increased the speed and accuracy of diagnosis, it is also enabling earlier detection of important diseases such as breast cancer. Although each of these technologies is independent, they are often employed together as part of an extended diagnostic workflow: medical imaging leads to a biopsy and examination of the biopsy results by a pathologist leads to a genomic study which is used to develop a treatment plan that is personalized to the patient's genome or observed genetic markers.

For each use case, this paper defines the technology, discusses how AI algorithms are being applied, and looks at the major challenges faced in research and clinical settings.

2 Medical Imaging

Medical imaging encompasses a wide range of modalities, from simple 2D x-rays to 3D CT and MRI scans to ultrasound. In most modalities, the goal is a visual representation of the body's interior that is used to identify damage or disease. These images are packed with an enormous amount of information and combing through them can be a challenge for even the most experienced professional.



As a discipline, medical imaging is under significant pressure to increase efficiency. Patient populations are aging, and experiencing conditions that require more imaging, but the size of the radiology workforce is flat or shrinking. Even in the first world, many countries have a shortage of radiologists, with particular shortages in rural areas. In the third world, lack of radiology expertise is widespread.

These shortages mean that radiology departments need to carry out more exams with fewer personnel. AI can serve as a valuable aid to radiologists, enabling departments to make better use of limited resources. By providing prescreening and preanalysis, AI can operate as a copilot, helping radiologists be more efficient and effective and flagging critical results—or exceptions—enabling radiology teams to focus their attention on patients in need of urgent care first.

Telemedicine is another area where AI-enabled screening is becoming valuable. In areas where there is shortage of expertise, or with large disabled or elderly populations, telemedicine can help bring the benefits of medical imaging to underserved populations.

2.1 AI in Medical Imaging

Machine learning methods are being applied in all areas of the medical imaging workflow, from image acquisition to analysis to reporting. For example, a medical startup company is developing a suite of medical imaging applications that can enable contrast and radiation dose reduction, up to 4 times faster scans, or both. As highlighted in a recent NVIDIA blog, this ability improves patient comfort and safety while increasing the productivity of a radiology department.

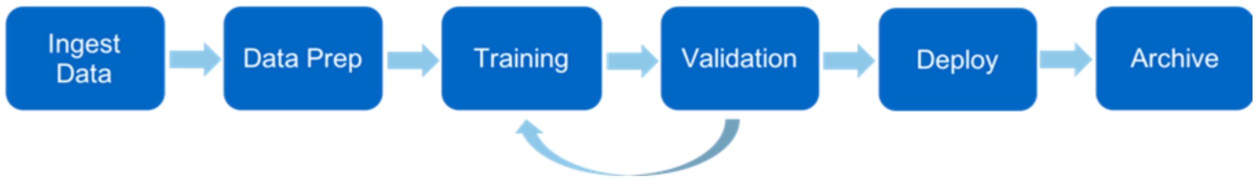
While AI efforts across the medical imaging workflow are important to overall progress, it is the use of AI for image analysis and diagnosis—computer-aided diagnosis—that has captured the most attention. Deep learning models are being developed for a wide range of conditions, promising to increase the speed and accuracy of analysis and enable earlier disease detection. High-profile areas under study include detection of lung nodules, brain cancer, multiple sclerosis, breast cancer, and prostate cancer.

AI is also being adopted to identify cardiovascular abnormalities and assess risks by measuring various heart structures. A visual assessment is often less accurate than an AI-powered system assessment. The orthopedics industry, in which hard-to-see fractures and other musculoskeletal injuries can lead to poor outcomes, is also adopting AI.

Some AI-powered diagnostic techniques are already moving beyond the clinic. The University of Michigan Kellogg Eye Center is combining a smartphone-mounted device for retinal imaging with a proprietary AI software platform called EyeArt. This solution can determine in real time whether a diabetic patient should see an ophthalmologist for follow-up.

Regardless of the disease under study, creating a deep learning model follows the general process illustrated in Figure 2.

Figure 2) Pipeline stages for deep learning.



The first step is to collect a dataset with examples of both diseased or damaged and healthy tissue for the target condition. The dataset must be prepared and, in most cases, clearly annotated. Today, annotation is usually a time-consuming manual process.

Much of the significant AI work to date has been accelerated through the use of publicly available, annotated datasets. For example, for lung nodule studies the Lung Image Database Consortium image collection (LIDC-IDRI) provides a set of computed tomography (CT) scans with annotated lesions. This dataset was used for the Lung Nodule Analysis 2016 (LUNA16) contest, which concluded in early 2018 with significant results.

The recipe for training deep learning models involves three ingredients:

- Search for the best model architecture
- Scale computation
- Accommodate large training datasets

Initial training can involve a lot of experimentation. Teams may try a variety of model architectures and starting parameters to achieve the best results. For medical image analysis (as with most image-oriented deep learning) convolutional neural network (CNN) architectures are typically employed. Popular models for medical image analysis include U-Net for 2D work and V-Net for 3D work, but many others exist.

The training process is compute intensive. Graphics processing units (GPUs) are used to parallelize computation and accelerate model training. Once a model has been trained, it is validated by using data separate from the training set. If the results of validation are satisfactory, a model can be deployed or distributed for wider testing. GPUs are also often needed to accelerate inferencing using the trained model.

The dataset used in training must be archived in such a way that it can be reused in the future.

2.2 Data Management in Medical Imaging

Starting about 25 years ago, medical imaging went through a process of digitization, so digital data management and standards are relatively mature. Most facilities use a commercial picture archiving and communication system (PACS) along with vendor-neutral archive (VNA) systems to store medical imaging data. Digital Imaging and Communications in Medicine (DICOM) is the established standard for managing and sharing medical image data.

For training, image files are typically copied from the PACS or VNA into a data lake or other external storage where it can be preprocessed. The total size of the training set depends on the modality or modalities used in training.

For example, in mammography a typical study has four images, each about 50MB in size, for a total of 200MB per mammogram. A training set based on 1,000 mammograms would be about 200GB of data. For scanning technologies like CT and MRI, each image is small, but there can be thousands of images in each study. A single MRI study can be multiple gigabytes, so training set sizes reach multiple terabytes quickly and memory capacity and I/O become a bottleneck. Although this is not an unmanageable amount of storage, it is something that has to be planned for, including consideration of the time it takes to move or copy data.

2.3 Research and Clinical Challenges in Medical Imaging

There are some significant AI challenges facing researchers and clinicians in medical imaging.

Research Challenges

For researchers, the biggest challenge is often sourcing enough annotated data, including examples of false positives. AI models learn by example; the more examples and the higher their quality, the better the results. However, consistently annotated data is uncommon, and data privacy concerns make patients and hospitals hesitant to share data, limiting training set size. One possible solution to this challenge is federated learning. The same AI model is trained serially in different locations to avoid moving data and to protect data privacy.

Another challenge that researchers face is gaining access to the necessary compute and storage resources. The cloud is attractive, especially for the early stages of AI exploration, assuming that patient privacy has been adequately protected and regulatory requirements satisfied. However, for large datasets, copying data into the cloud over a network connection may not be possible, and cloud storage costs can mount quickly.

Clinical Challenges

In clinical settings, a major concern is how to incorporate diagnostic AI models into the medical imaging workflow. In addition to numerous university and medical center research teams, there are hundreds of startup companies doing AI work for medical imaging. As a result, there are potentially dozens of AI models that could be deployed clinically, each with its own infrastructure needs and workflow requirements.

Because of the urgency that is often associated with the analysis of medical images, using the cloud may not be an option in clinical settings, both because of the time needed to move data to the cloud for analysis and the risk of loss of connectivity. It's likely that hospitals will want the necessary high-speed compute and data storage to support AI inferencing on their premises.

By investing in the right infrastructure, a hospital or medical center can be positioned to support dozens of different models without making additional hardware investments for each, potentially including AI for medical imaging, computational pathology, genomics, and other areas.

A final problem in clinical settings is specificity. In order to get the greatest accuracy from a trained model, it may be desirable to fine tune a model using annotated data from an institution's own patients and instruments. Again, this model requires GPUs and high-speed data storage, either on the premises or in the cloud, as well as staff with the necessary experience.

3 Digital Pathology

The pathology workforce is facing a situation very similar to that in medical imaging—demand for pathology services is growing faster than the number of pathologists. This means that pathology labs have to become more efficient in order to handle more cases in less time.

In traditional pathology, slides are prepared from patient tissue samples and then reviewed by a pathologist under a microscope at high magnification. This manual process can be error prone and time consuming, especially if the pathologist needs to consult with an outside expert.

This is where digital pathology offers assistance. Once tissue samples are prepared, the resulting slides are scanned using whole slide imaging (WSI). A pathologist can then review slide images on a full-color computer monitor rather than looking at the slides directly through a microscope. Pathologists can share images and collaborate with a few mouse clicks, easily perform critical measurements on screen, and compare patient images with reference images.

Pathology labs are also increasingly able to use computational pathology, applying a variety of numerical and machine learning techniques to digital slide images to increase the speed, accuracy, and efficiency of diagnosis.

3.1 AI in Computational Pathology

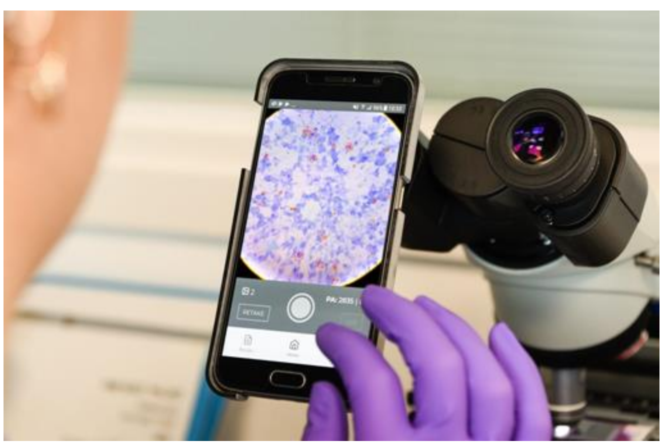
Although computational pathology is similar in many ways to AI in medical imaging, there are some significant differences. In general, digital pathology is several years behind medical imaging in terms of AI maturity.

In part, this may be because pathology has been relatively slow to digitize. In medical imaging, digitization offered a clear path to cost reduction and increased workflow efficiency. However, digital pathology adds digital technology on top of existing physical processes, so cost benefits are less clear. Physical slides must still be prepared and stained, and they must be archived for extended periods. Perhaps as a result, fewer established, annotated digital pathology datasets exist for AI experimentation.

The size and complexity of digital pathology images creates an additional challenge. A single case may involve multiple images totaling from 0.5 to 6GB in size, depending on the level of magnification. A single image may be thousands of pixels in length and width, and captured in full color. Rather than having deep learning algorithms operate on a whole image, it's common to divide the image into patches and analyze each patch, adding significantly to the workload for both training and inference.

With these differences accounted for, training proceeds in a manner similar to that for medical imaging, with the goal of training a CNN to accurately distinguish between disease conditions and normal cell tissue. For example, the digital pathology startup Proscia has created a CNN model, called DermAI, specifically for analyzing skin lesions. Proscia claims that the model can automatically classify hundreds of variants of skin diseases. Many other startups are focusing on various aspects of digital pathology.

BacilAi combines deep learning with low-cost hardware for the treatment of tuberculosis. TB is the second largest cause of death by infectious disease in the developing world, and its high mortality rate is in large part due to a lack of available, affordable diagnosis and inconsistent results. BacilAi uses a smartphone to capture images from an ordinary lab-grade microscope. The system analyzes sputum images using a deep learning algorithm to identify, count, and classify TB cells to determine the patient's disease state.



3.2 Data Management in Digital Pathology

Given its relative youth, digital pathology lacks the mature data management environment of medical imaging. A standard PACS system can't accommodate digital pathology images. Each WSI scanner may have a proprietary image format. However, digital pathology images can be managed in PACS-like systems that are now becoming available, and an extension to the DICOM standard has been defined to encompass digital pathology.

3.3 Research and Clinical Challenges in Digital Pathology

The much larger file sizes—and resulting dataset sizes—of digital pathology are likely to be a significant challenge in both research and clinical settings.

Research Challenges

Digital pathology is well established in both nonprofit and for-profit research settings. Although annotated data remains a challenge for AI research, larger institutions have the resources and partnerships to overcome these limitations.

Given the size of individual imaging files and resulting training datasets, infrastructure is a much bigger challenge for AI projects in digital pathology than for medical imaging, in terms of both compute and data storage. Development work may begin in the cloud, but more mature teams that have big datasets and a clear roadmap need on-premises resources.

Clinical Challenges

In clinical settings, digital pathology faces many of the same challenges as medical imaging. With many deep learning algorithms available, each trained for a specific use case, it will be difficult to determine how to meaningfully incorporate AI into a standard pathology workflow. The size of the data files, combined with the fact that multiple slides are typically associated with a single patient, makes it highly unlikely that labs will be able to use cloud resources for inferencing. Medium to large labs will need high-speed compute and data storage to support AI inferencing on the premises.

4 Genomics

The first human genome took more than 10 years to sequence at a cost of billions of dollars. Today, the same task can be completed in a day or two for a thousand dollars. As a result, the amount of genomic data, human and otherwise, is exploding, and genomics—the study of genes, their functions, and their interactions—is rapidly becoming a clinical tool that can be used alone or in conjunction with medical imaging and digital pathology. Combining medical imaging and/or digital pathology with genomics allows doctors to understand how a patient's genes (genotype) manifest physically in the patient (phenotype).

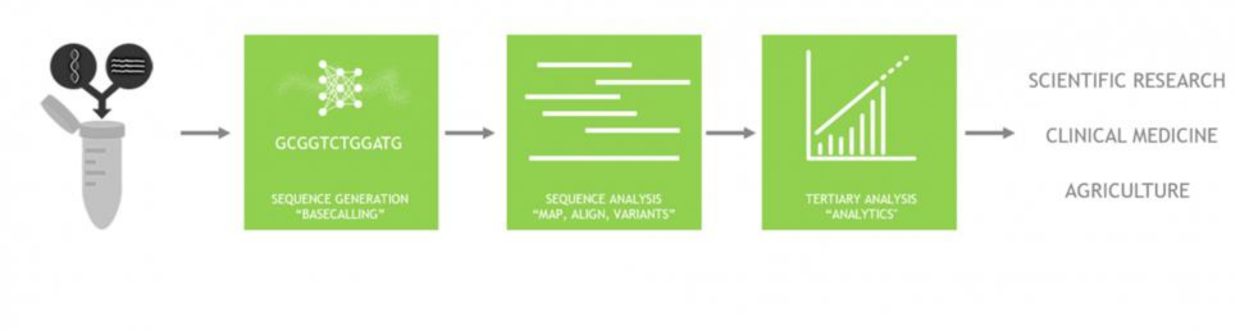
Genomics is essential to deliver on the promise of precision medicine. However, success increasingly requires machine learning to make sense of the enormous volumes of genetic data and to identify clinically significant correlations and patterns quickly as an aid to diagnosis and treatment.

4.1 AI in Genomics

The fundamental challenge of genomics is to take mountains of human sequence data and figure out which differences are important. Which gene variants, or combinations of genes, contribute to various medical conditions, and how do you use genomic information to individualize patient treatment?

NetApp customer WuXi NextCODE has created a unique platform designed specifically to organize, mine, share, and apply genomic data to improve human health. Over the past 20 years, it has amassed the world's largest database of human genome sequences. Through AI, WuXi is bringing genomics studies from the realm of theoretical benefits to the real world, in the service of real patients.

Figure 3) Whole genome sequencing.



Whole genome sequencing (WGS) and analysis is valuable in clinical settings only if it can be done quickly, accurately, and inexpensively. In many cases only limited genetic data is available in hospitals and medical labs because of the time and cost to process and store WGS data. Because the amount of data generated per patient can be 300GB to 1TB, processing alone can take several days.

WGS is a three-phase process, with AI playing a role in each phase of the analysis. During sequence generation, automated sequencers employ numeric and machine learning techniques to optimize output. The output data coming from sequencers goes immediately to a secondary analysis to map and align the data and identify where the sequence varies from a reference sequence—a process referred to as variant calling.

NetApp and NVIDIA Parabricks has developed high-performance GPU computing and deep-learning technologies to accelerate these operations, providing 30 to 50 times faster secondary analysis compared with CPU-based approaches.

Once a patient's genome has been sequenced and the sequence has been analyzed, a tertiary analysis is necessary to determine which of the identified variants may have clinical significance. Open source tools such as Exomiser and commercial products from companies such as Diploid and BC Platforms are being developed to help with various aspects of the medical genomics challenge, including tertiary analysis. Tertiary results can guide diagnosis and treatment, but they have to be presented in a way that clinicians can understand and act on.

Once a diagnosis has been made, pharmacogenomics can also be applied to help determine the most appropriate treatment based on the patient's genetic makeup.

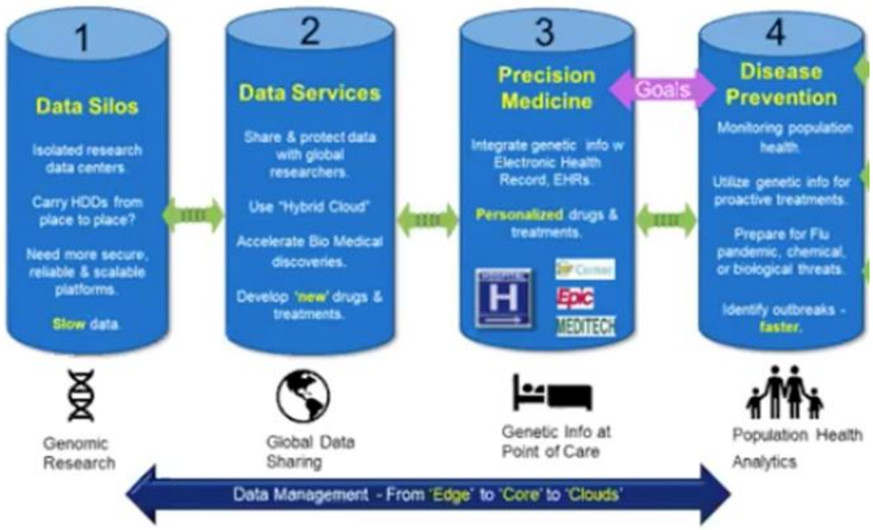
As an example of what is becoming possible, San Diego researchers recently set a record for sequencing to diagnosis in a neonatal/pediatric ICU. The team created a pipelined workflow that included Illumina sequencers and Diploid MOON to automatically filter and rank the possible causative gene variants, compressing the entire workflow down to about 19 hours.

4.2 Data Management in Genomics

Data management is a much more significant challenge in genomics than in medical imaging or digital pathology. With sequencing results for a single individual ranging up to 1TB in size, WGS creates data management challenges in both research and clinical settings. While the file formats used in genomics are standardized, there is no equivalent to a PACS or VNA for managing sequence data. Methods to store genomic data in EHR systems are being investigated, but existing EHR databases aren't well suited to storing these large data files directly. It’s likely that EHR systems will need to store pointers to the data files in an external archive optimized for the purpose.

IT maturity for genomics data management proceeds from silos of data in isolated research centers to a data services model with shared data access. High maturity will be needed to enable precision medicine at the point of care as well as population health and disease prevention.

Figure 4) Data management stages for genomics.



4.3 Research and Clinical Challenges in Genomics

As described earlier, data management is a fundamental challenge for genomics researchers. In many big data applications, data loses its value over time. In genomics, the data never loses its value. Intermediate data generated during analysis is frequently used for reanalysis, enabling new scientific discoveries. Data scientists in pharmaceutical companies frequently reanalyze genomic files to try to discover new mutations or biomarkers. This reanalysis poses a new set of scalability challenges.

Reanalysis can generate duplicate files. Then there are clinical reports and reference datasets that need to be carefully versioned and deduplicated. Researchers need solutions that can effectively deduplicate and compress data. Special compression techniques offered through technical partners such as PetaGene can shrink BAM and FASTQ file types used in genomics by 10 times.

Many researchers are turning to the cloud to gain access to necessary storage and compute resources. Because of the large datasets and the amount of compute required, genetics researchers such as WuXi NEXTcode use a hybrid cloud approach to gain access to the necessary storage and compute resources. Data from WuXi’s on-premises sequencers is uploaded to the cloud for analysis and study.

Another NetApp customer uses genomics to guide the development of cell-based treatments for a variety of cancer types. The team relies on FlexPod® AI and NetApp FAS storage on their premises in combination with Cloud Volumes ONTAP® in AWS for additional capacity to accommodate rapid growth.

In clinical settings, secondary and tertiary analysis of patient genomic data will have to be done close to the sequencer, at least in urgent cases. Hospitals and laboratories that want to perform sequencing for clinical use will need to have local compute and high-performance storage, and the results of genetic analysis must be integrated with EHRs so that they are quickly available to clinicians.

5 NetApp and NVIDIA Solutions for AI in Healthcare

Table 1) Comparison of AI in medical imaging, digital pathology, and genomics.

	Data per Patient (Relative)	ML Algorithms	Research	Clinical
Medical Imaging	Low Up to 1GB	CNN	Algorithm Training	Fine-tuning Inferencing
Digital Pathology	Medium Up to 6GB	CNN	Algorithm Training	Inferencing
Genomics	High Up to 1TB	CNN, RNN, other	Comparative Studies Algorithm Training	Inferencing

This white paper covers three different use cases for AI in healthcare. Although there are significant differences and unique challenges, there are also commonalities that can be used to guide decision making for research and clinical organizations that want to prepare for AI.

NetApp and NVIDIA are partnering to deliver AI solutions for the healthcare industry. Both companies are laser-focused on eliminating AI bottlenecks and advancing the realm of the possible at a rapid pace. NetApp's attention to the data pipeline amplifies NVIDIA's efforts to accelerate compute.

By combining technologies from both companies, ONTAP AI accelerates all facets of AI training and inference to deliver better outcomes more quickly. The solution brings together NVIDIA DGX servers, NetApp cloud-connected all-flash storage, and Cisco Nexus or Mellanox Spectrum switches. This proven architecture simplifies, integrates, and accelerates both machine learning and deep learning algorithms, allowing customers to start small and grow as needed without disruption. The ONTAP AI Toolkit offers an array of tools and functions to simplify setup and operation, delivering immediate productivity.

NVIDIA and NetApp are also building an ecosystem of AI partners that can help solve healthcare AI challenges and accelerate success. Many of the companies mentioned in this white paper are NVIDIA Inception Program members, NetApp AI Partner Network members, or both.

Where to Find Additional Information

To learn more about the information that is described in this document, review the following websites:

- AI and Analytics for Healthcare (NetApp)
  https://www.netapp.com/us/artificial-intelligence/healthcare-ai-analytics/index.aspx
- Healthcare and Life Sciences (NVIDIA)
  https://www.nvidia.com/en-us/industries/healthcare-life-sciences/

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

WP-7314-1219