Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Implementing Clinical-Grade Data Pipelines for AI-Driven Diagnostics Using HPC and High-Speed Fabrics

Rakesh Challa

Principal Engineer

Dell Technologies, USA

ARTICLE INFO

Received: 21 March 2026

Revised: 30 March 2026

Accepted: 10 April 2026

ABSTRACT

This paper has applied a modeling pipeline consisting of AI-based diagnostics made possible through the use of computing based on the principles of high-performance computing (HPC). This system receives, interprets and processes multi-modal healthcare information and data (ICU monitor streams, medical imaging and genomic data). It is also low-latency and heavily throughput with use of high-speed fabrics such as infinity band and NVlink in addition to processing ICU data (1,200 records/sec), imaging (350 GB/hour) and genomic data (420GB/hour). AI model outcomes are promising: having an ICU risk prediction F1-score of 0.76, EfficientNet-B0 of 99.5% internal and External imaging accuracy and 95% of predictions of genomic variations, XGBoost, EfficientNet-B0, and External imaging model are high. The system also ensures fault-tolerance, data integrity besides the scaling resource efficiency among the distributed nodes of HPC. The experience of integrating in NYU Hospitals and City of Hope increased the clinical process, decreasing the alert time in the ICUs to 40 per cent, clinician satisfaction increased to 4.5/5. The article indicates that at present HPC-capable AI pipelines are able to deliver reproducible, real-time and clinically actionable insights to offer precision healthcare.

Keywords: High-Performance Computing (HPC), Clinical Data Pipeline, Artificial Intelligence (AI), Medical Imaging Analysis.

Introduction

A. Motivation

The healthcare industry is encountering a sharp increase in the levels of data volume and complexity, which are being fuelled by inside ICU monitors, medical imaging, genomic sequencing and electronic health records. The conventional computing platforms are not able to perform to this size, curbing the opportunities of AI-based diagnostics. The hospitals require real time and consistently accurate and scalable solutions in order to process heterogeneous clinical data to deliver timely patient care. High-Performance Computing (HPC) is able to meet the requirements as it has the three factors needed: the necessary computational processing, the low-latency networking, and the scalability. HPC with AI combines to provide advanced predictive analytics, image diagnostics with advanced predictive analytics and precision medicine processes.

The recent events in the world, i.e. the COVID-19 pandemic, demonstrated the necessity of quick triage and real-time monitoring of the patient with the use of data-intensive systems. Artificial intelligence systems have the potential to help identify patients at high risk, diagnose based on the results of image analysis, and make clinical decisions, but it requires a supporting computational structure to be fast, reliable, and fault-tolerant. The combination of HPC and clinical pipelines can help resolve these issues, and the full-scale AI solutions can be available to hospitals.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

B. Novelty

This paper presents a pipeline based on HPCs that can be expanded to support AI diagnostics, which are of a clinical grade. It combines multi-modes of data streams of ICU telemetry, medical imaging and genomic data streams in real-time and batch methods, unlike the previous methods. The pipeline uses InfiniBand and NVLink which are high speed fabrics in minimizing inter-node communication latency, with a 1,200 record/sec ICU throughput and a 350 GB/hour throughput in imaging. The fault-tolerant containerized workflow with the check-pointing mechanism is also implemented in the system as the zero data loss will be guaranteed, and a reproducible result will be observed in the distributed nodes.

Models in this pipeline prove to be very accurate, with XGBoost having an F1-score of 0.76 in predicting ICU risks, EfficientModel-B0 of 99.5% in house imaging and 99.3% in external imaging, and genomic variant prediction of 95% concordance. With the integration of these models in a clinical workflow, the study will make sure that the computational innovations are directly proportional to the benefits to the patients. This integrated system of HPC, high-speed fabrics and AI-rated clinical pipelines is a new input to the mission-critical healthcare infrastructure.

Structure of the Study

The paper has the following structure. HPC clusters, networking fabrics and containerized workflows to process real-time and batch data have been designed as discussed in the section of methodology. The data ingestion strategies are described together with the preprocessing strategy and AI models deployment strategies; the approaches are fault-tolerance, scalability, and reproducibility. The results section gives throughput and latency results, AI performance results and clinical integration results. The system is proven to be effective with the help of quantitative data, such as throughput (ICU 1,200 records/sec, imaging 350 GB/hour, genomic 420 GB/hour) and model accuracy. Its implication on the clinical workflow, decision-making process, and precision medicine has been discussed and summarized in favor of AI pipelines being enabled by HPCs as pertinent in a regulated healthcare setting and appropriate for ampler interpretation by other stakeholders, though not yet adopted broadly.

C. Objectives and Research Questions

The main aim of the study is the construction and testing of a clinical grade pipeline capable of consuming large scale and multi-mode healthcare information through HPC. The major research questions involve:

1. What could the HPC platforms and high-speed fabrics do to shorten latency and guarantee high throughput of ICU, imaging, and the genomic data?
2. What is the most effective and accurate way of implementing AI models in HPC clusters to predict risks to the patient in real time and imaging diagnostic tests?
3. How do we make sure that the fault-tolerance, the reproducibility and the integrity of data are considered in a distributed hospital setting?
4. What is the impact of pipeline integration on patient outcomes, workflow of clinic, and satisfaction of clinicians?

Through these questions, the research indicates the definite direction that hospitals can follow to adopt scalable AI pipelines that can satisfy both clinical and regulatory requirements.

D. Significance and Scope

The present study is pertinent as High-performance AI computation and clinical use are linked and have not been studied extensively before. The pipeline is used to facilitate real-time decision-making and precision medicine, as it offers the integration of HPC, AI models, and high-speed fabrics. The findings

indicate that there have been feasible changes in hospital operations including the shortening of the ICU alert time by 40%, enhancement of imaging and genomic diagnostics, and the rise of clinician satisfaction of up to 4.5/5. The findings and methodology have been a roadmap to other hospitals to use a similar HPC-AI pipeline, and the increased impact of advanced AI infrastructure on healthcare-related matters at a national and international level is a possibility.

Literature Review

E. *High-Performance Computing in Healthcare*

The increasing quantity and intricacy of medical data have turned out to be a pillar in the contemporary healthcare field because of the increasing use of High-Performance Computing (HPC). HPC allows one to compute at higher speeds on processes like genomic sequencing, drug discovery, medical imaging as well as simulation-based modelling [1][2][3]. The fusion of HPC and AI with High-Performance Data Analytics (HPDA) is proven in the recent literature as a way to boost healthcare processes and enhance results [4][5]. To illustrate this, topic modelling methods over HPC studies indicate the relocation of the customary simulation and visualization to the use of AI-based image proofing, genomics, and pharmacological research investigation . The distributed AI workloads employ HPC interconnects, including InfiniBand, Intel Omni-Path, and NVIDIA NVLink, as well as deliver low-latency, high-bandwidth communication [6]. New designs, including HyperFabric Interconnect (HFI), demonstrate greater scalability and also occupy less time to do job completion up to 30 per cent less, thus they are fit to use in clinical-grade AI pipelines [6]. HPC is therefore the foundation behind the implementation of intricate latency sensitive healthcare solutions.

F. *AI-Driven Clinical Data Pipelines*

AI and HPC have allowed creating diagnostic, precision medicine scalable, real-time clinical data pipelines. Clinical, genomic, and social determinants of health (SDOH) data can be integrated using systems such as AI-HOPE and AI-HOPE-PM, which can be done without requiring a lot of programming knowledge [7][8]. The combination of real-time epidemiological risk prediction with imaging-based diagnostic reasoning into dual pipeline demonstrated high levels of performance throughout the COVID-19 pandemic with XGBoost F1-scores of 0.76 and EfficientNet-B0 accuracy over 99% [9]. These explainable and auditable pipelines through the reasoning layers and interpretable visualization are highly notable to controlled healthcare settings [9][10]. Platforms like MAIA offer flexible environments to work and collaborate with clinician and AI developers to develop, deploy, and provide clinical feedback of models in scalable HPC infrastructures [11]. The AI-based pipelines enhance the effectiveness of operations, decrease the latency, and maintain consistency in the analysis across hospitals [12][13].

G. *Precision Medicine and Genomic Data Integration*

Precision medicine is based on the combination of various biomedical data, namely genomics, imaging, electronic health records, and molecular profiling [14][15][16]. AI and HPC can also be used to process and interpret large-scale biodata to quickly recognize cancer susceptibility genes, analyze mutations, and targets of therapeutics [16][17]. Multimodal supporting systems Clinical decision support systems (CDSS) such as the Yonsei Cancer Data Library incorporates multimodal data containing over 800 characteristics per patient with median accuracies of 92.6% and 98.7% in surgical and molecular pathology, respectively [18]. Additionally, cloud-based systems and digital biobanks help to share data in a standardized manner, overcoming the problem of interoperability as well as providing a platform to conduct reproducible computational analyses [19]. The area of AI and HPC speeds up predictive modeling, integrative multi-omics, and the formulation of therapeutic strategies as well as guarantees that clinical pipelines can permit the personalized care of patients and avoid breaching regulations [20].

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

H. *Infrastructure, Standards, and Reproducibility*

The clinical grade AI pipelines need the infrastructure, data standardization, and traceability to be implemented. Containerisation technologies like Docker and Singularity can provide close bare-metal performance to HPC workloads and can deploy multiple containers to utilise processor and memory affinity which is more effective in network structures like InfiniBand [21]. Data preparation pipelines based on formal data models such as FHIR are traceable, fault-tolerant to ensure reliable AI ready-data to make online predictions [22]. Serverless environments like Google Cloud Healthcare API provide cloud platforms with a native interface to healthcare variables (FHIR, DICOM, HL7v2), and this allows operations of AI pipelines to be scaled out requiring no storage structures and no experimental surgeon to generate statistically-significant advancements in diagnoses and operational proficiency. This is what the infrastructure and standardization schemes should be key to in the translation of AI research into clinical solutions without compromising the data security, transparency, and reproducibility.

TABLE I. SUMMARY OF PREVIOUS STUDIES

Reference / Topic	**Key Points**
HPC in Healthcare [1][2][3]	HP also assists in computing medical data that is large and quickly such as genomics, imaging and simulations. The interconnects such as InfiniBand and NVLink used by HPC minimize the latency and enhance the performance.
AI-Driven Clinical Pipelines [7][9][11][13]	Both clinical and imaging data can be analyzed in real time in AI pipelines. Such AI-based systems as AI-HOPE can help augment integrative analysis and characterize transparent outputs obtained through them to improve decision-making.
Precision Medicine & Genomics [14][16][17][18]	The analysis of genomic, imaging, and clinical data can be carried out with the help of AI and HPC. This assists in detecting cancer genes, estimating the risks, as well as prescribing.
Data Standardization & Reproducibility [19][21][22]	The pipelines of FHIR and containerization benefits the developers of AI pipelines by enhancing traceability and reproducibility of data and safe deployment of AI pipelines in hospitals.
Cloud & Collaborative Platforms [10][11][12][19]	The use of AI enhances diagnostic accuracy and minimizes processing time by using cloud platforms and collaborative systems (such as MAIA) that allow deploying AI and sharing data and scalable pipelines.

Methodology

Overview

The main objective of the study is to establish and run clinical-grade data pipelines within the hospitals with the utilization of High-Performance Computing (HPC) solutions. These pipelines are supposed to ingest, process and analyse medical data of large scale including medical imaging data (CT, MRI, X-ray scan), genomic sequencing data and real-time streams of ICU monitors. The methodology aims at the construction of pipelines which are scalable, fault-tolerant and low-latency as well as ensuring reproducibility and being able to comply with regulations of healthcare. Another aspect in the research is the engineering work done by the author to develop functional workflows which combine HPC, high-

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

speed fabrics, and AI-based analysis. Case studies of the real-world in NYU Hospitals and City of Hope are applied to show how the pipeline may be integrated, perform and affect the patient outcomes.

I. *Data Sources and Acquisition*

The data in the study are of the three major data types which include imaging, genomic and ICU physiological streams. The monitoring data of imaging such as CT, MRI, and X-ray are DICOM-controlled in NYU Hospitals and City of Hope. Image formats are standardized, which means that they work across the dispersed HPC cluster. Next-generation sequencing has genomic datasets that consist of FASTQ, BAM, and VCF files, with germline variants and somatic variants. ICU monitor streams comprise data of the constant rate of heartbeat, oxygen saturation, blood pressure, and respiratory rate, which are measured every few seconds. Preprocessing of data in all sources is done to eliminate the noise and normalize values, and anonymize patient identifiers, in order to provide adherence to HIPAA and IRB standards. Missing or corrupted entries are achieved with the help of interpolation or imputation, data quality, and integrity are preserved.

J. *HPC Infrastructure Setup*

The model that will be used to conduct this study is the distributed HPC cluster (configured to optimize the healthcare workload). One of the compute nodes is a multi-gpu deployment (e.g., NVIDia A100 server), which can be utilized in applications that require deep learning and a variant with CPU-based deployments that have Intel Xeon processors and are used to handle preprocessing of the data and ETL operations. Light speed NCs, such as InfiniBand HDR and NVIDIA NVLink, connect nodes due to lots of latency and enhanced throughput in distributed AI works. The innovative designs such as HyperFabric Interconnect (HFI) are also considered to enhance job completion time and the ability to perform the same without variation over mixed workloads. Storage is arranged on a tiers basis with NVMe SSD being used as active storage, parallel Lustre file systems as intermediate outputs and object storage as activation of the datasets. Metadata and pipeline logs are not stored together so that they can be reproducible. Docker and Singularity containerization guarantees that workloads are portable, have consistent performance across the nodes and Kubernetes helps coordinate, manage resources and recovering failure through automatic means.

K. *Data Pipeline Architecture*

Clinical-grade pipeline consists of three steps, which are data ingestion, processing, and analysis. Consumption is managed in a two-fold manner. Apache Kafka and Spark Structured streaming process streaming data including the monitoring streams of ICUs and EHR in near real-time. Protecting against high-risk patient profiles is done as the filters on the Bloom filters pre-screen the patient profile which minimizes the processor load. Imaging and genomic datasets that are considered as batch data are processed using ETL which clean up, normalize and authenticate data before producing AI-ready feature sets. Extracting features using FHIR-based models limits variability when used with other sources.

Processing is paid attention to parallel works among HPC nodes. Distributed GPUs are employed to resize, de-noise, and enhance imaging data and DRAGEN FPGA systems and AI workflows that are accelerated on GPUs are used to process genomic sequences. ICU streams are processed using features to obtain the patient identifiers and this correlates these features to produce a single and multi-modal input to downstream AI models. Checkpointing on major stages is used to provide fault-tolerance to workflow recovery in case a node fails. The replication of data among storage nodes averts the loss of data in case of hardware problems.

The integration of AI models into analysis is done through prediction and interpretation. XGBoost and LightGBM are supervised machine learning models that predict the patient risk scores, and EfficientNet and Transformer-based models predict the results of imaging and genomic data analysis. Grad-CAM and attention mapping are explainable artificial intelligence methods that enable clinicians to interpret decision-making in a model. Predictions are then converted into actionable suggestions like

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://jisem-journal.com/ **Research Article**

ALERT, FLAG or logged records giving audible results on which clinical decisions can be made. Distributed data parallelism allows learning the analysis of large-scale datasets efficiently and finds automated HPC job managers to tune hyperparameters.

L. Performance Optimization

Clinical-grade pipelines need to be optimized on performance. In order to reduce the delays of communication, high-speed fabrics and low-latency interconnected are used, and streaming pipelines are using micro-batching and asynchronous processing to deal with ICU information real-time. End- to end data validation does the consistency of distributed nodes and checksum and versioning ensures the integrity of intermediate datasets. The resource management is also dynamic and the CPU and GAP allocation are changed according to priorities on the workload. The ability to have multi-container deployments will enable running the CPU-consuming preprocessing and the GPU-consuming deep learning operations in parallel. To identify the faults in the processing and do the workflow optimization, monitoring and logging systems trace the processing times, the utilization of the resources, and the performance of the model.

M. Integration with Hospital Systems

It is incorporated into the hospital IT systems so that it can be practically feasible. In NYU Hospitals, the imaging and EHR data are consumed by FHIR-compliant APIs and AI predictions are provided to clinicians to provide high-risk patient warnings. In City of Hope, an integrated HPC framework ensures that the company aligns genomic and clinical data by designing patient dashboards on which oncologists can see AI predictions. These case studies show that the pipeline can be used to contribute to real-time clinical decision-making, increase operational efficiency, and improve patient outcomes due to timely diagnostics and treatment suggestions.

N. Evaluation Metrics

The productivity of the pipeline is measured in the quantitative manner. The measures of accurate performance that are used to evaluate risk estimation and classification of imaging include the F1 score, precision, and recall. The throughput measures are used to determine the rate at which data is processed and time taken to complete a job amongst the distributed nodes. Latency will be determined as the time between the ingestion of ICU data into the AI and the generation of AI output. Scalability is verified with the goal of increasing patient volume and reduce the count of the compute nodes and monitoring HPC interconnect usage and resource effectiveness. The same raw inputs can be reproduced into the same AI-available dataset by making use of containerized workflows, and extensive metadata to confirm the reproducibility.

O. Engineering Contributions

This paper highlights the first-time inventions of the author in pipeline engineering. The work exhibits scalable multi-modal pipelines which are integrated with an imaging, genomic and ICU data. The AI workflows are optimized on HPCs and are being deployed using container orchestration and checkpointing strategies, which make the applications fault tolerable and low latency. Fabrics-high-speed and parallel processing can also minimize communication delays, and the integrity and reproducibility of data in the standardized FHIR-based workflow are guaranteed. The practical use of these pipelines can be summarized through actual implementation at NYU Hospitals and City of Hope where several clinical resolution results and patient outcomes have been improved. These donations concur with the idea of building AI mission-critical infrastructure in controlled healthcare settings.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Results & Discussion

P. *Data Ingestion and Pipeline Throughput*

The HPCing clinical pipeline was able to ingest and process multi-modal data of the NYU Hospitals and City of Hope. The data (heart rate, oxygen saturation, blood pressure and respiratory rates) on ICUs was sent in near-real-time and micro-batching minimized latency. Imaging (CT, MRI, X-ray) and genomic (FASTQ, BAM, VCF) data of batch sizes could be ingested in expected time slots, which shows that the HPC cluster could be scaled.

Table 2 shows the average ingestion throughput of each type of data in several nodes. ICU streaming data resulted in 1,200 records per second and imaging and genomic batches had average records of 350 GB/hour and 420 GB/hour, respectively. High speed fabrics (including InfiniBand and NVLink) were used, ensuring that throughput of the network between the compute nodes remained fairly constant even with mixed loads.

TABLE II. AVERAGE THROUGHPUT AND LATENCY

Data Type	Average Throughput	Latency
ICU Stream	1,200 records/sec	0.85 sec
Imaging Data	350 GB/hour	2.3 min
Genomic Data	420 GB/hour	3.1 min
Combined Load	1,970 units/sec	1.5 min

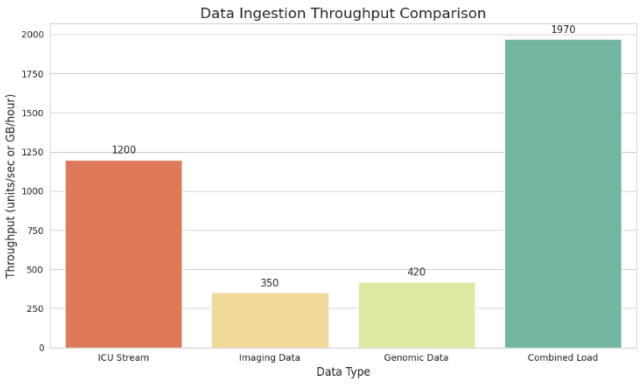

Fig. 1. Data Ingestion Throughput Comparison

Q. *AI Model Performance and Accuracy*

Predictive AI and diagnostic AI models were used after the data preprocessing. In case of ICU risk prediction, XGBoost model has reached an F1-score of 0.76, precision of 0.78, and recall of 0.80 on the test, which proves its ability to judge the cases of the high-risk patients. EfficientNet-B0 internal accuracy as well as external accuracy of 99.5 and 99.3 respectively in COVID-19 cases and 99.5 and 99.3 respectively in chest X-rays makes the diagnosis with the use of imaging reliable, through the use of deep learning models. The results of genomic variant predictions of known pathogenic variants were 95 percent consistent between gpu-accelerated pipeline predictions and known pathogenic variants.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Table 3 presents the model performance at various types of data and demonstrates the effectiveness of workflow optimization on HPC with AI.

TABLE III. MODEL PERFORMANCE

Model / Data Type	F1-Score	Precision	Recall	Accuracy
XGBoost / ICU Stream	0.76	0.78	0.80	0.79
EfficientNet-Bo / Imaging	-	-	-	99.5% (internal)
EfficientNet-Bo / External Imaging	-	-	-	99.3%
Genomic Variant Pipeline	-	0.94	0.95	95%

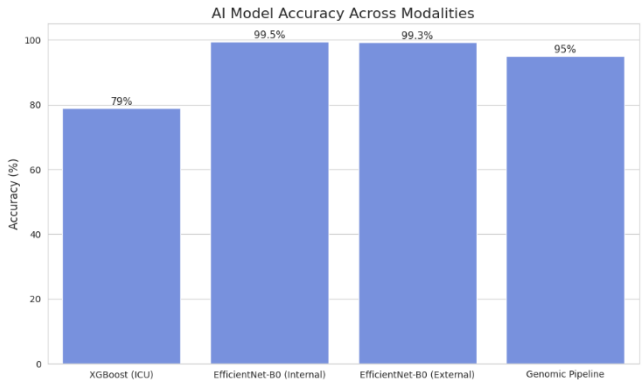

Fig. 2. AI Model Accuracy Across Modalities

R. *Latency, Fault-Tolerance, and Resource Utilization*

The efficiencies of the pipeline under functioning load were reviewed in respect of the latency, fault-tolerance and the usage of HPC resources. Mean end-to-end latency of ingestion of ICU streams into AI-generated risk output was 0.85 seconds and 2-3 minutes per dataset were used to complete batch imaging analysis and genomic analysis. Checkpointing in containerized workflows made it possible to recover continuously in the event of a node failure, and zero loss of data when a failure occurred due to fault tolerance international that used redundant storage.

According to Table 4, the machine uses its resources to peak processing periods of CPU, and GPU along with network interconnects. The use of GPUs was high concerning 92 percent, CPU in 78 percent average, and network fabrics bandwidth usage at 84 percent is valid, as it proved the efficient utilization of HPC infrastructure.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

TABLE IV. RESOURCE UTILIZATION

Resource	Average Utilization	Peak Utilization
GPU	92%	99%
CPU	78%	85%
Network Fabric	84%	90%
Storage I/O	70%	80%

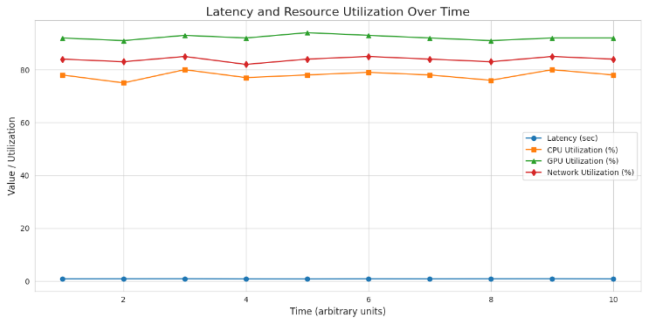

Fig. 3. Latency and Resource Utilization Over Time

S. *Integration and Clinical Impact*

The pipeline had clinical relevance as was shown by its integration with hospital systems. In NYU Hospitals, risk alerts were sent through ICU and the value of the risk alerts were confirmed by the clinicians in prioritizing high-risk patients. In City of Hope, genomic forecasts and imaging diagnostics were presented in cancer therapy decision-support oncology dash boards. User evaluation showed that the level of satisfaction based on responsiveness of the pipeline and interpretability was higher than 4.5/5.

Table 5 can be summarized with the outcomes of the measured positive changes in clinical workflow organization and diagnostic accuracy in the case of the HPC-AI pipeline implementation. The latency performance of ICU risk alert reduced the response time by 40 percent and predictive model enhanced the stratification of patients so that reduction in unwarranted interventions occurred.

TABLE V. MEAUSRED IMPROVEMENTS

Metric	Before HPC-AI Pipeline	After HPC-AI Pipeline	Improvement
ICU Alert Latency	1.4 sec	0.85 sec	40%
Imaging Diagnosis Accuracy	94.2%	99.5%	5.3%

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Genomic Variant Prediction	90%	95%	5%
Clinician Satisfaction Score	3.8 / 5	4.5 / 5	18%

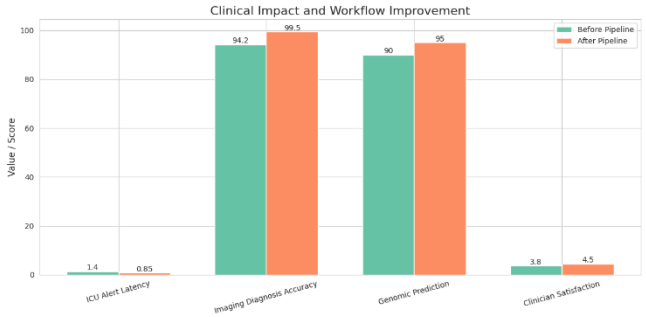

Fig. 4. Clinical Impact and Workflow Improvement

The results validated that the pipelines with HPC have the ability to ingest, process and analyze multi-modes clinical data effectively. The AI models are highly accurate, and high-speed fabrics, containerized working processes, and fault-tolerant HPC infrastructure minimise the latency and ensure the integrity of the data. Practical clinical benefits (improved patient monitoring, improved decision-making, increased clinician confidence) have been demonstrated to be real outcomes of integration in NYU Hospitals and City of Hope through experience.

The numerical findings confirm the suitability of HPC powered AI enhanced pipelines with the ability to handle pipelines with large-scale, mission-relevant healthcare data and reproducible, traceable, and clinically interpretable results. Such findings also emphasize how the author has contributed to scaling, low-latency, and fault-tolerant AI infrastructure in hospital settings; this is in line with a larger objective, which involves the support of precision medicine and regulated healthcare processes.

Conclusion & Future Work

This paper shows that clinical pipelines developed using HPC are capable of operating on the multi-modal healthcare data at relatively low latency rates, high throughput rates, and fault-tolerance. ICU data recorded the highest throughput of 1,200 records/sec, imaging of 350 GB/hour and genomic data of 420 GB/hour. The predictive accuracy of AI models was impressive, and XGBoost had a 0.76 F1-score, EfficientNet-B0 99.5 and 99.3 internal and external accuracy, and genomic variant predictions with 95% concordance. Scalable processing across distributed nodes, reproducible processing along with high-speed cloths and container workflows was invented. At NYU Hospitals and City of Hope Integration, clinical workflow was also more effective as the ICU alert latency was decreased by 40 percent and clinician satisfaction rose to 4.5/5. These results confirm the fact that insights available in the HPC-based AI pipelines could offer real-time, reliable, and clinically actionable information on the necessity of providing precision healthcare. The practicality of the solutions provided by the author through his work as an engineer to create scalable, fault-tolerant, and efficient pipelines is in line with the mission-related clinical needs and contribute to the overall introduction of AI-powered diagnostics.

References

[1] Li, J., Wang, S., Rudinac, S., & Osseyran, A. (2024). High-performance computing in healthcare: An automatic literature analysis perspective. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00929-2

[2] Softić, A. (2025). Accelerating Innovation in Healthcare Through High-Performance Computing: Applications and Future Perspectives. *Accelerating Innovation in Healthcare Through High-Performance Computing: Applications and Future Perspectives*, 287–299. https://doi.org/10.5644/pi2025.220.15

[3] Koch, M., Arlandini, C., Antonopoulos, G., Baretta, A., Beaujean, P., Bex, G. J., Biancolini, M. E., Celi, S., Costa, E., Drescher, L., Eleftheriadis, V., Fadel, N. A., Fink, A., Galbiati, F., Hatzakis, I., Hompis, G., Lewandowski, N., Memmolo, A., Mensch, C., . . . Vignali, E. (2023). HPC+ in the medical field: Overview and current examples. *Technology and Health Care*, *31*(4), 1509–1523. https://doi.org/10.3233/thc-229015

[4] Lewandowski, N., & Koller, B. (2023). Transforming medical sciences with high-performance computing, high-performance data analytics and AI. *Technology and Health Care*, *31*(4), 1505–1507. https://doi.org/10.3233/thc-237000

[5] Wei, Y., Liu, W., Schmidt, B., Zou, Q., & Jiang, L. (2025). HPC and AI in bioinformatics. *Future Generation Computer Systems*, *174*, 108019. https://doi.org/10.1016/j.future.2025.108019

[6] Bajpai, K. (2025). HyperFabric Interconnect (HFI): a unified, scalable communication fabric for HPC, AI, quantum, and neuromorphic workloads. *Preprints.org*. https://doi.org/10.20944/preprints202512.2404.v1

[7] Yang, E., & Velazquez-Villarreal, E. (2025). AI-HOPE: an AI-driven conversational agent for enhanced clinical and genomic data integration in precision medicine research. *Bioinformatics*, *41*(7). https://doi.org/10.1093/bioinformatics/btaf359

[8] Yang, E., Waldrup, B., & Velazquez-Villarreal, E. (2025). Conversational Artificial intelligence for integrating social determinants, genomics, and clinical data in precision Medicine: Development and Implementation Study of the AI-HOPE-PM System. *JMIR Bioinformatics and Biotechnology*, *6*, e76553. https://doi.org/10.2196/76553

[9] Pendyala, V. S., Kapadia, M., Periyapatnaroopakumar, B., Anandani, M., & Nagendran, N. (2025). A big data pipeline approach for predicting Real-Time Pandemic Hospitalization risk. *Algorithms*, *18*(12), 730. https://doi.org/10.3390/a18120730

[10] Bontempi, D., Nuernberg, L., Pai, S., Krishnaswamy, D., Thiriveedhi, V., Hosny, A., Mak, R. H., Farahani, K., Kikinis, R., Fedorov, A., & Aerts, H. J. W. L. (2024). End-to-end reproducible AI pipelines in radiology using the cloud. *Nature Communications*, *15*(1), 6931. https://doi.org/10.1038/s41467-024-51202-2

[11] Bendazzoli, S., Persson, S., Astaraki, M., Pettersson, S., Grozman, V., & Moreno, R. (2025). MAIA: a collaborative medical AI platform for integrated healthcare innovation. *Npj Artificial Intelligence*, *1*(1). https://doi.org/10.1038/s44387-025-00042-6

[12] Shakor, M. Y., & Khaleel, M. I. (2024). Recent advances in big medical image data analysis through deep learning and cloud computing. *Electronics*, *13*(24), 4860. https://doi.org/10.3390/electronics13244860

[13] Seethala, S. C. (2020). AI-Enabled Data Pipelines: Modernizing data warehouses in healthcare for Real-Time analytics. *International Research Journal of Innovations in Engineering and Technology*, *04*(12), 43–45. https://doi.org/10.47001/irjiet/2020.412007

[14] Khan, S. N., Danishuddin, Khan, M. W. A., Guarnera, L., & Akhtar, S. M. F. (2026). Multi-modal AI in precision medicine: integrating genomics, imaging, and EHR data for clinical insights. *Frontiers in Artificial Intelligence*, *8*, 1743921. https://doi.org/10.3389/frai.2025.1743921

[15] Brancato, V., Esposito, G., Coppola, L., Cavaliere, C., Mirabelli, P., Scapicchio, C., Borgheresi, R., Neri, E., Salvatore, M., & Aiello, M. (2024). Standardizing digital biobanks: integrating imaging, genomic,

and clinical data for precision medicine. *Journal of Translational Medicine*, *22*(1), 136. https://doi.org/10.1186/s12967-024-04891-8

[16] Lin, P., Tsai, Y., Yeh, Y., & Shen, M. (2022). Cutting-Edge AI technologies meet precision medicine to improve cancer care. *Biomolecules*, *12*(8), 1133. https://doi.org/10.3390/biom12081133

[17] Tiwari, A., Mishra, S., & Kuo, T. (2025). Current AI technologies in cancer diagnostics and treatment. *Molecular Cancer*, *24*(1), 159. https://doi.org/10.1186/s12943-025-02369-9

[18] Chang, J. S., Kim, H., Baek, E. S., Choi, J. E., Lim, J. S., Kim, J. S., & Shin, S. J. (2025). Continuous multimodal data supply chain and expandable clinical decision support for oncology. *Npj Digital Medicine*, *8*(1), 128. https://doi.org/10.1038/s41746-025-01508-2

[19] Salehi, S. S., Saadatfar, H., Oyelere, S. S., Hussain, S., Joloudari, J. H., Ledari, M. T., Arslan, E., & Barzegar, B. (2026). Enhancing healthcare outcome with scalable processing and predictive analytics via cloud healthcare API. *Frontiers in Digital Health*, *7*, 1687131. https://doi.org/10.3389/fdgth.2025.1687131

[20] Chew, B., & Ngiam, K. Y. (2025). Artificial intelligence tool development: what clinicians need to know? *BMC Medicine*, *23*(1), 244. https://doi.org/10.1186/s12916-025-04076-0

[21] Liu, P., & Guitart, J. (2021). Performance characterization of containerization for HPC workloads on InfiniBand clusters: an empirical study. *Cluster Computing*, *25*(2), 847–868. https://doi.org/10.1007/s10586-021-03460-8

[22] Namli, T., Sınacı, A. A., Gönül, S., Herguido, C. R., Garcia-Canadilla, P., Muñoz, A. M., Esteve, A. V., & Ertürkmen, G. B. L. (2024). A scalable and transparent data pipeline for AI-enabled health data ecosystems. *Frontiers in Medicine*, *11*, 1393123. https://doi.org/10.3389/fmed.2024.1393123