Research Article

Novel Hybrid Edge-Cloud Framework for Efficient and Sustainable Omics Data Management

Rani Adam,[1] Daniel R. Catchpoole,[2,3,4] Simeon S. Simoff,[1] Paul J. Kennedy,[5] Quang Vinh Nguyen[1]

[1]School of Computer, Data & Mathematical Sciences, Western Sydney University, Sydney, Australia
[2]The Tumour Bank, Children's Cancer Research Unit, Kids Research, The Children's Hospital at Westmead, Sydney, Australia
[3]The Discipline of Paediatrics and Child Health, The Faculty of Medicine, The University of Sydney, Sydney, Australia
[4]Faculty of Information Technology, The University of Technology Sydney, City, Australia
[5]Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia

Address correspondence to Rani Adam (22104050@student.westernsydney.edu.au).

Sources of Support: None. Conflict of Interest: None.

Submitted: Jun 27, 2024; First Revision Received: Oct 22, 2024; Accepted: Oct 25, 2024; Published online: Dec 27, 2024

Adam R, Catchpoole DR, Simoff SS, Kennedy PJ, Nguyen QV. Novel hybrid edge-cloud framework for efficient and sustainable omics data management. *Innov Dig Health Diagn Bio*. 2024; 4:81–88. DOI: 10.36401/IDDB-24-5.

ABSTRACT

Introduction: The healthcare landscape is rapidly evolving through the integration of diverse data sources such as electronic health records, omics, and genomic data into patient profiles, enhancing personalized medicine and system interoperability. However, this transformation faces challenges in data integration and analysis, compounded by technologic advancements and the increasing volume of health data. **Methods:** This study introduces a novel hybrid edge-cloud framework designed to manage the surge of multidimensional genomic and omics data in the healthcare sector. It combines the localized processing capabilities of edge computing with the scalable resources of cloud computing. Evaluations involved using simulated cytometry datasets to demonstrate the architecture's effectiveness. **Results:** The implementation of the hybrid edge-cloud framework demonstrated improvements in key performance metrics. Network efficiency was enhanced by reducing data transfer latency through localized edge processing. Operational costs were minimized using advanced compression techniques, with the Zstandard (ZSTD) codec significantly reducing data size and improving upload times. The framework also ensured enhanced data privacy by leveraging edge-based anonymization techniques, which process sensitive information locally before transfer to the cloud. These findings highlight the framework's ability to optimize large-scale omics data management through innovative approaches, achieving significant gains in scalability and security. **Conclusion:** Integrating edge computing into a cloud-based omics data management framework significantly enhances processing efficiency, reduces data size, and speeds up upload times. This approach offers a transformative potential for omics and genomic data processing in healthcare, with a balanced emphasis on efficiency, cost, and privacy.

Keywords: omics data management, genomic data management, health data integration, hybrid edge-cloud framework, health data edge processing

INTRODUCTION

The healthcare landscape is rapidly transforming owing to technologic advancements and the growing availability of health data, with an emphasis on integrating diverse data sources like electronic health records, genomic sequencing, and wearables into comprehensive patient profiles to improve care and facilitate evidence-based decisions.[1–5] This integration is crucial for personalized medicine, public health management, and enhancing interoperability across healthcare systems. However, realizing the full potential of health data integration presents several challenges. The multidimensional and complex nature of health data poses significant integration and analysis hurdles. Compounded by the rapid pace of technologic advancement in healthcare, the generation of new data types is accelerating. The volume of stored data has grown by 40 PB between 2016 and 2018[6] alone, outpacing Moore's law, which demonstrates an exponential increase rather than a linear trend. This rapid data growth, combined with the increasing diversity of data types, underscores the critical need for efficient data integration, consolidation, and management. The current landscape of health data tools is mostly characterized by a dichotomy between on-premises and cloud-based solutions, each presenting unique advantages and challenges in terms of control, cost, and convenience.

On-premises or local solutions such as AbioTrans,[7] compcodeR,[8] CANEapp,[9] BioAnalyzer,[10] and GNomEx[11] specifically address the intricate needs of genomic research and other omics data analysis. These platforms excel in omics analysis, alternative splicing, and single-cell RNA sequencing, among other functionalities, offering a blend of gene expression analysis, statistical distribution fitting, and correlation analysis. They present advantages in terms of predictable costs, enhanced data privacy, and faster data access speeds. However, they face limitations in scalability and collaboration and incur increased operational costs, posing challenges for large omics and genomic dataset analysis.[12]

In contrast, cloud-based architectures herald a new era of accessibility and scalability characterized by both horizontal scaling by adding more virtual machines (VMs) of the same size and type as needed to increase parallelism and vertical scaling. Cloud-based architectures enhance the computational capabilities of existing VMs by upgrading their size/type within a series/family or switching to a VM series/family with superior capabilities, such as those based on newer central processing units.[13] In health data analysis, the integration of cloud computing with distributed computing and machine learning has introduced innovative approaches to managing large-scale bioinformatics data.[14,15] Platforms such as the Cancer Genomics Cloud[16] and G-DOC Plus demonstrate the potential of cloud-based solutions in merging genomic and multi-omics data with electronic health record information and managing diverse biomedical data, respectively. The versatility of these solutions is further illustrated by OncDRS[17] and Spark-Seq,[18] which facilitate precision medicine applications and the analysis of next-generation sequencing data.

Despite their substantial contributions, these frameworks require further exploration and enhancement to fully address scalability, collaboration, and the efficient management of large genomic and omics data volumes. Existing systems struggle to keep pace with the massive data volumes produced by next-generation sequencing techniques.[19] This issue is compounded by the dependency on consistent and reliable internet access. In regions where internet access is not universally available, this dependency could potentially limit the effectiveness and reach of cloud-based solutions.[12] The costs associated with cloud storage, while eliminating some upfront expenses, can lead to higher long-term costs due to ongoing subscription fees.[20] Furthermore, the regulatory landscape in healthcare poses considerable challenges. There are significant concerns regarding data security and the necessity to comply with various healthcare regulations.[21] Ensuring that cloud computing solutions adhere to these legal and ethical standards is essential for their effective and responsible use.

This article introduces a novel hybrid edge-cloud framework that addresses scalability and adaptability. Tailored specifically for genomic and other omics applications, our framework optimizes data processing for real-time and near real-time analytics by strategically combining the strengths of edge and cloud computing. Edge computing refers to the practice of processing data near the source of its generation rather than in a centralized data-processing warehouse, improving response times and saving bandwidth.[22] The focus on omics data is justified by its complexity, volume, and importance in personalized medicine, as it includes genomics, transcriptomics, proteomics, and metabolomics, all crucial for understanding disease mechanisms and developing targeted therapies. Its high dimensionality and privacy concerns present challenges for processing and management. To address these challenges, the framework emphasizes network efficiency, operational cost reduction, and data privacy, which are critical for managing large-scale omics data securely and efficiently. The specific objectives of this study are:

- Present a novel framework that uniquely integrates edge computing optimized for near real-time omics data analytics processing with edge-based anonymization techniques, such as SHA-256 hashing,[23] to ensure data security before transmission to cloud storage. This dual approach significantly enhances network efficiency, operational cost-effectiveness, and data protection.
- Use state-of-the-art compression codecs like Zstandard (ZSTD), a near real-time compression algorithm developed by Facebook, using LZ77 combined with fast Finite State Entropy and Huffman coding,[24] to reduce data storage and transfer costs while saving time.
- Improve of scalability and accessibility by processing data at the edge.

Network efficiency, operational cost reduction, and data privacy were selected as primary outcomes owing to their importance in securely and cost-effectively managing high-volume omics data. These metrics are critical for evaluating the practical feasibility and adoption of the framework in clinical and research settings. This study hypothesizes that the hybrid edge-cloud framework will (1) improve network efficiency for omics data processing, (2) reduce operational costs through effective data compression, and (3) enhance data privacy compliance by minimizing data exposure during processing.

METHODS

Our hybrid edge-cloud framework is designed for omics applications, leveraging established methodologies from traditional frameworks, as proposed by Mrozek [13] and Nguyen et al.[25] It uses edge-cloud computing to optimize

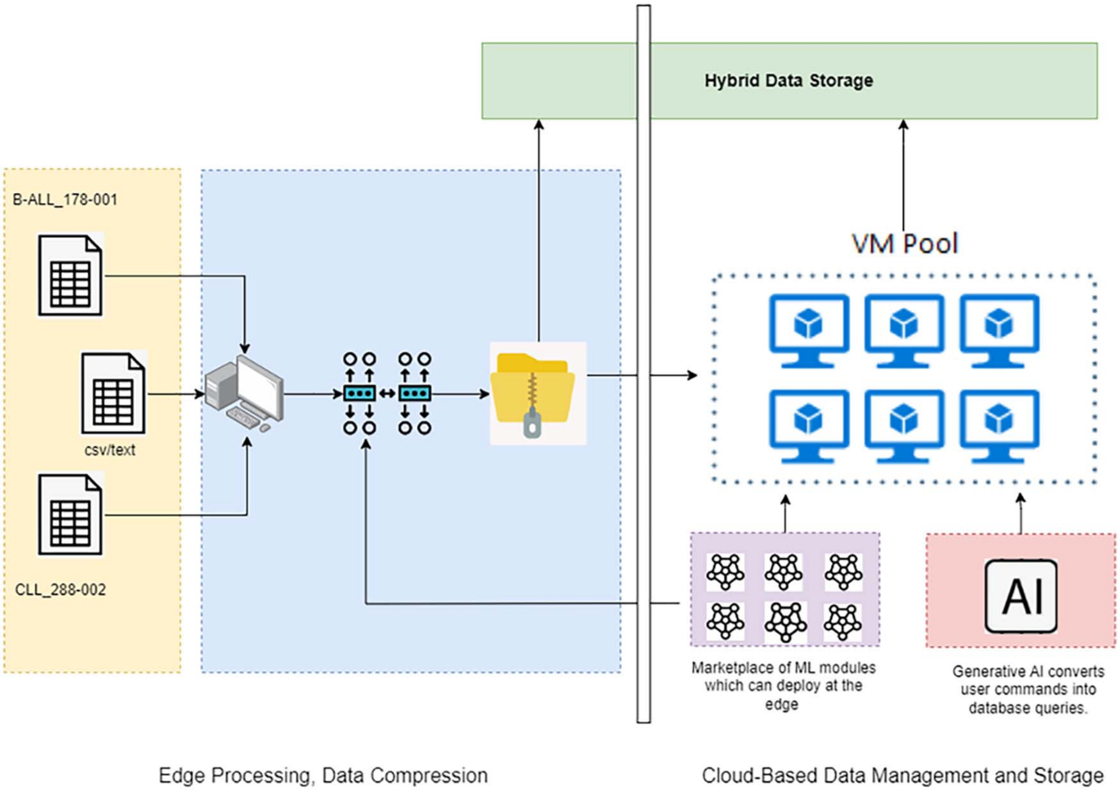
Hybrid Data Storage
B-ALL_178-001
csv/text
CLL_288-002
VM Pool
AI
Marketplace of ML modules which can deploy at the edge
Generative AI converts user commands into database queries.
Edge Processing, Data Compression
Cloud-Based Data Management and Storage

Figure 1. The proposed hybrid edge-cloud framework architecture. AI: artificial intelligence; B-ALL: B-cell acute lymphoblastic leukemia; CLL: chronic lymphocytic leukemia; CSV: comma-separated values; ML: machine learning; VM: virtual machine.

data processing near the source, as shown in Figure 1. This approach minimizes response times and conserves bandwidth.[22] Furthermore, Figure 2 visually illustrates the main components and workflow of the hybrid edge-cloud framework. It provides a detailed graphical representation of the interactions between edge computing processes, cloud storage solutions, and data management techniques.

To evaluate the framework, we used both simulated and real cytometry datasets. Formal ethical approval was not required for this phase of the project as our focus was on developing systems architecture rather than analyzing personal or sensitive information. Furthermore, all data were de-identified, ensuring that there was no risk of re-identification.

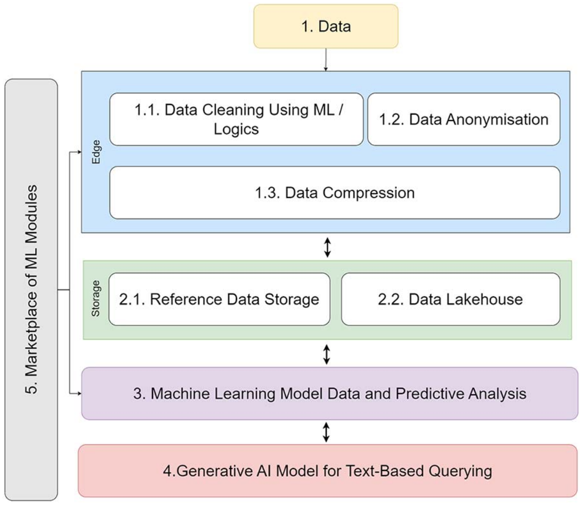
1. Data
5. Marketplace of ML Modules
Edge
1.1. Data Cleaning Using ML / Logics
1.2. Data Anonymisation
1.3. Data Compression
Storage
2.1. Reference Data Storage
2.2. Data Lakehouse
3. Machine Learning Model Data and Predictive Analysis
4.Generative AI Model for Text-Based Querying

Figure 2. The proposed hybrid edge-cloud framework workflow. AI: artificial intelligence; ML: machine learning.

Dataset Specifications

- **Simulated dataset:** The simulated dataset mimics chronic lymphocytic leukemia (CLL) samples, including both CLL cases and normal controls. It encompasses measurements of 10 cellular markers and six scatter parameters crucial for cytometry analysis. The dataset varies in size, ranging from 50,000 to 10 million rows. This variety allows for rigorous testing under diverse load conditions.
- **Real patient data:** This study also used real patient cytometry data from B-cell acute lymphoblastic leukemia (B-ALL) and CLL to assess the framework's effectiveness and computational efficiency:
  - **Dataset 1 (B-ALL):** Contains 6.3 million records (approximately 1.64 GB), featuring a broad spectrum of 11 cellular markers. This extensive data volume and marker variety provided a rich environment for rigorous analysis and performance testing.
  - **Dataset 2 (CLL):** Similar to dataset 1 but larger, with 12 markers and 41 million records (approximately 2.07 GB). Its massive scale offers an exceptional opportunity

to evaluate the framework's scalability and efficiency, particularly in edge-based cytometry data processing and analysis.

Implementation Environment

In this research, initial data processing within edge computing environments was performed using a standard laptop equipped with an 11th Gen Intel Core i7-1185G7 processor, 16.0 GB of RAM, and an internet connection speed of 20 Mbps. The operating system used was Windows 10 Pro (64-bit), and all algorithms were implemented using Python (version 3.8), with relevant packages such as ZSTD (version 0.23.0), SNAPPY (version 0.7.3), LZ4 (version 4.3.2), and GZIP (part of the standard Python library). The cloud component used Microsoft Azure services, integrating Azure Data Lakehouse for data storage. This setup ensures consistency and replicability across different environments.

Edge Computing and Initial Data Processing

Initial data processing within edge computing environments involved the following steps:

- **Data cleaning (Figure 2, Step 1.1):** This critical phase starts with validating the accuracy and completeness of the dataset. It includes identifying and flagging duplicate entries for removal and pinpointing irrelevant information. A key technique used during this phase is the use of variance threshold–based selection criteria. This approach helps in systematically identifying data that exhibit minimal variation and are thus considered redundant or irrelevant for the analysis. The identification and flagging process can be significantly automated and enhanced through advanced machine learning models. An exemplary tool in this regard is the FlowClean algorithm,[26] available as a package for the R programming language, which can be installed from the machine learning models marketplace.
- **Data anonymization (Figure 2, Step 1.2):** To ensure privacy and security, data anonymization SHA-256 hashing[23] techniques are used to convert personal identifiers and sensitive data into unique hash values, effectively masking the original data while preserving its utility for analysis. Additionally, tokenization methods are used to replace sensitive data elements with nonsensitive equivalents, known as tokens, which can be mapped back to the original data only through a secured tokenization system. To enhance data privacy compliance, tokens matching anonymized data can be securely exported and stored locally on physical storage devices, remaining outside of cloud systems. This setup adheres to country-specific privacy laws and provides an additional layer of security, limiting access only to authorized personnel. These anonymization practices adhere to stringent data protection standards, including the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the ISO/IEC 27001:2013 international standard for information security management. By conforming to these standards, organizations ensure the confidentiality, integrity, and security of sensitive information, safeguarding against unauthorized access and breaches.
- **Data compression and upload (Figure 2, Step 1.3):** This model of the framework is optimized for preparing data for storage and access. The principal technique used in the module is an advanced compression strategy that incorporates both the Parquet file format and the ZSTD compression algorithms. The Parquet format excels in managing large datasets due to its efficient columnar storage capability, which significantly minimizes storage space while maximizing read and write speeds. Concurrently, ZSTD provides high compression rates,[27] further reducing the data footprint in storage. These technologies together aim to maximize storage efficiency and boost the speed of data retrieval, key factors for scalable and cost-effective data management.

Cloud-Based Data Management Storage

Data management and storage are facilitated by using Azure and Azure Data Lakehouse. This approach integrates two key components:

- **Relational database (Figure 2, Step 2.1):** Microsoft SQL Server 2022 databases are used for metadata management, reference data, schema, and access policies. This ensures structured storage, efficient data retrieval, and robust schema enforcement.
- **Data Lakehouse (Figure 2, Step 2.2):** The Data Lakehouse used for storing compressed data can handle large volumes of unstructured omics data. The scalability and support for diverse data ingestion are critical for managing complex datasets.

To ensure data availability and resilience, the cloud storage infrastructure uses geo-replication, automatically replicating data across multiple geographically distinct data centers. This approach ensures that even if one region experiences a failure, the system can seamlessly retrieve data from another location, maintaining uninterrupted access. In conjunction with edge devices' local buffering, geo-replication enhances disaster recovery capabilities and minimizes data loss risks during cloud outages.

Testing Strategies

To evaluate our proposed edge-computing framework, we conducted two distinct case studies. The first used a simulated dataset with CLL characteristics, and the second leveraged real patient cytometry data from both B-ALL and CLL to assess the framework's performance under varied complexities.

Both case studies focused on assessing the effectiveness of data compression, speed, and cost at the edge layer within the hybrid framework. The experimental setups for both studies were standardized to ensure consistency in testing conditions. Across the case studies, we compared the performance of four leading compression codecs—SNAPPY, LZ4, ZSTD, and GZIP—on key metrics relevant to edge computing, such as compression ratio, network bandwidth utilization, and storage cost impact.

Data Analysis

This study used quantitative methods to analyze the efficiency of different compression codecs, specifically focusing on their ability to reduce data size and upload times. The evaluated codecs included ZSTD, Snappy, LZ4, and GZIP, selected for their known efficiency in data compression. The evaluation involved comparing these codecs, based on their compression ratios and speed of execution, aligning with the study's objectives of reducing storage costs and optimizing upload times. The analysis calculated the percentage reduction in data size and the improvements in upload times after compression for each codec.

Other Design Components

- **Open platform marketplace (Figure 2, Step 5):** This component is envisioned as an innovative open marketplace designed to foster collaboration among researchers and developers by enabling the seamless exchange and application of machine learning models. Modelled after GitHub, but specifically tailored for scientific applications, this marketplace offers flexibility in infrastructure management. Users can choose a centrally managed option, maintained by a scientific consortium or academic institution, ensuring governance, sustainability, and compliance with privacy regulations. Alternatively, users can opt for a self-hosted version, allowing organizations to retain full control over their infrastructure. A key advantage of this setup is that it eliminates the need for data to be transferred to external servers, thereby significantly enhancing data privacy and reducing bandwidth usage. Additionally, models could be deployed directly at the network's edge, which would optimize data cleaning and analysis processes. This component is not evaluated. It is part of the framework and workflow process, serving as a conceptual foundation for future development and implementation. By providing flexibility and leveraging a familiar platform structure such as GitHub, the marketplace supports diverse needs within the scientific community and minimizes costs.
- **AI-driven querying engine (Figure 2, Step 4):** The AI-driven querying engine represents a conceptual leap in data retrieval technology, using generative AI to interpret natural language queries and convert them into precise database operations; this innovative engine would enable users to access data simply by describing their needs in plain language, eliminating the necessity for complex query syntax. This approach significantly reduces the learning curve and speeds up the adoption process, making data access more democratic, intuitive, and user-friendly. Although promising, this component is not evaluated. It is part of the framework and workflow process, serving as a conceptual foundation for future development and implementation.

Reasoning for Method Selection

These methods were chosen over alternative approaches owing to their compatibility with the high dimensionality of omics data and the scalability required for near real-time edge computing. Purely on-premises or entirely cloud-based solutions were found insufficient for balancing scalability, security, and cost-effectiveness. In contrast, our hybrid edge-computing approach offers significant advantages for real-time processing by leveraging both local and cloud resources efficiently. This study adheres to the STROBE reporting guidelines,[28] as outlined by the EQUATOR network, ensuring transparency and rigor in the methodologic process.

RESULTS

The study analyzed data from two distinct datasets: a simulated dataset mimicking CLL characteristics and real patient cytometry data. The CLL dataset comprised simulations with parameters for 10 cellular markers and six scatter parameters, and the real patient data included cytometry results from 41 million records covering both patients with B-ALL and CLL.

Both case studies demonstrated that ZSTD outperforms other codecs in terms of data reduction, achieving an average compression rate of 56.09% compared to GZIP's 51.32%, Snappy's 20.23%, and LZ4's 23.78%. These average compression rates were calculated by summing the percentage compression achieved for each file within the datasets and dividing by the total number of files, providing a representative value for each codec's performance. The analysis yielded effect sizes of $d = 0.95$ for ZSTD compared to GZIP, $d = 7.17$ for ZSTD compared to Snappy, and $d = 6.46$ for ZSTD compared to LZ4, indicating moderate to large improvements over GZIP and substantial improvements over Snappy and LZ4. All comparisons showed statistically significant differences ($p < 0.05$, 95% CI). This establishes ZSTD as the most effective codec, significantly reducing storage requirements. For example, it compressed a dataset of 47.3 million records from 3.71 GB to 1.60 GB, achieving an average data size reduction of 56.87%. This reduction translates to monthly cost savings based on Azure LRS Hot storage pricing, which ranges from $1.70 to $2.08 per GB per month. Figures 3 and 4 visually summarize our findings on compression efficiency for case studies 1 and 2. Additionally, our analysis revealed that using ZSTD resulted in a 57% reduction in data upload times due to its compression efficiency.

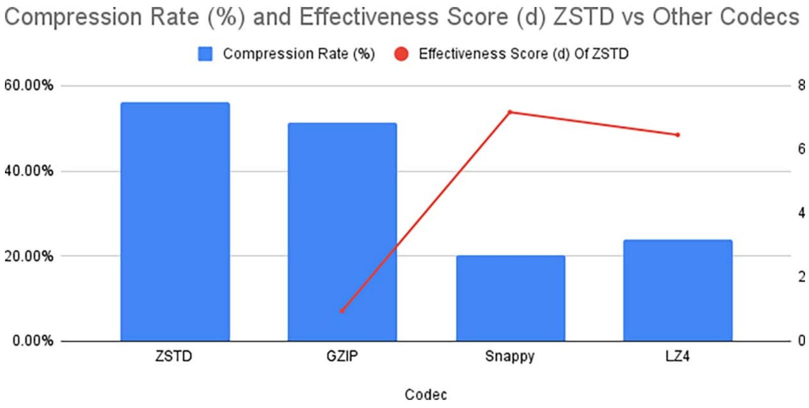

Figure 3. Compression rate (%) and effectiveness score (d) of ZSTD compared to other codecs. ZSTD: Zstandard.

To calculate the effect size (Cohen d), the following formula was used:

$$d = \frac{M_1 - M_2}{SD_{pooled}},$$

where M_1 and M_2 are the mean compression rates for ZSTD and the comparison codec (GZIP, Snappy, and LZ4); SD_{pooled} is the pooled standard deviation of the two groups, calculated as:

$$SD_{pooled} = \sqrt{\frac{(n_1 - 1)SD1^2 + (n_2 - 1)SD2^2}{n_1 + n_2 - 2}},$$

where $(n_1 - 1)SD1^2 + (n_2 - 1)SD2^2$ are the sample sizes for each group; $SD1^2$ and $SD2^2$ are the standard deviations for each group.

By applying this formula, we demonstrated that ZSTD achieves a statistically significant effect size of d = 0.95, supporting its superior performance in data reduction. Assuming a dataset representative of large-scale genomic projects, such as 100 million genomic reads per sample and 1000 samples, the original data size could reach approximately 40 TB. This aligns with typical sizes for whole genome sequencing (100–200 GB per sample) as reported by projects like the 1000 Genomes Project[29] and The Cancer Genome Atlas.[30] Implementing ZSTD would reduce storage needs to about 16 TB, resulting in estimated monthly cost savings of \$41,779.20 to \$51,148.80 in a cloud setting with standard pricing. This broader context underscores the potential impact of using ZSTD for large-scale data management in genomic research.

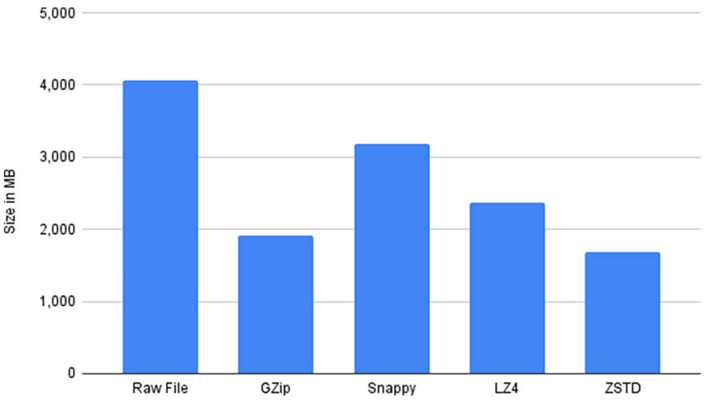

Figure 4. The results of the compression using GZIP, SNAPPY, LZ4, and ZSTD. ZSTD: Zstandard.

DISCUSSION

This study evaluated the efficacy of a hybrid edge-cloud framework in managing and analyzing cytometry data, with a particular focus on CLL and B-ALL. The experimental findings demonstrated that our framework, especially when using the ZSTD codec, substantially improved data compression and reduced upload times, affirming our hypothesis that edge computing can enhance data processing efficiency in genomic studies.

Our integration of edge computing within a hybrid framework for cytometry data is an area that has not been extensively explored in existing literature. The significant gains in data compression and upload speed are likely a result of processing data close to its generation sites, thereby reducing the latency and bandwidth usage often associated with centralized cloud processing. We demonstrated a capable framework that not only mitigates bandwidth and storage costs but also enhances the privacy and security of sensitive health data. Our findings emphasize the framework's dual capacity to lower operational costs and safeguard sensitive health information, making it a compelling solution for handling complex datasets in high-stakes environments.

Our results align with advancements in cloud-based genomic solutions, such as the Cancer Genomics Cloud,[16] which similarly manage and analyze diverse data types. However, our hybrid framework differs significantly by introducing an edge component, enhancing local data processing capabilities, particularly in environments with intermittent or unreliable internet access. This contrasts with studies that focus solely on cloud-based processing, where upload times and compression efficiency may be affected by network latency and bandwidth limitations. By integrating edge computing, we reduce the dependency on stable internet connectivity, offering improved upload times and compression efficiency when compared to purely cloud-based models. This suggests that while existing cloud-based solutions achieve moderate efficiency, our edge-cloud hybrid model provides superior performance, especially in clinical settings with limited infrastructure. Furthermore, although other frameworks have adopted cloud-based models, few have embraced the hybrid approach because of infrastructure complexities and regulatory constraints. By addressing these challenges, our framework demonstrates its versatility in managing complex health data and improving operational efficiency without relying exclusively on cloud connectivity and without sacraficing compliance or performance.

Although ZSTD proved most effective for cytometry data compression, its efficacy may vary with other data types. As different genomic and clinical data formats emerge, exploring tailored codec options optimized for specific data types may be beneficial. Future research could help establish guidelines for codec selection, aligning data types with specific compression techniques to standardize and enhance processing efficiencies.

Although the proposed framework shows significant promise, several limitations must be acknowledged. Firstly, the reliance on cytometry datasets may not fully capture the complexities and variances encountered in genomics and omics data. Additionally, the initial setup and configuration of the hybrid edge-cloud infrastructure can be resource-intensive and may pose challenges for smaller organizations. The framework's performance and scalability need further validation through real-world implementations and diverse datasets to ensure its robustness across different use cases. Another limitation is that the envisioned marketplace component for machine learning models and the AI-querying engine component have not been evaluated within this study.

Real-world implementation of the framework may encounter challenges such as the initial setup's resource intensity, the need for technical expertise in edge-cloud integration, and ensuring compliance with privacy regulations like General Data Protection Regulation (GDPR) and HIPAA. For practical adoption, a gradual rollout in low-stakes environments is recommended to refine the configuration process while addressing privacy, security, and integration concerns. Such an approach allows for controlled testing and adjustments, ensuring that the framework meets regulatory standards and integrates smoothly with existing healthcare infrastructure.

Future work will focus first on optimizing data-cleaning techniques and refining compression processes at the edge to enhance real-time processing capabilities and reduce computational overhead. Following this, the development of a dynamic marketplace for machine learning models will be prioritized, allowing researchers to collaborate by submitting, evaluating, and using models, thereby fostering innovation and increasing the framework's adaptability across various genomics projects. Additionally, we plan to explore dynamic data compression strategies tailored to diverse genomic and omics data types, which will enhance processing flexibility and efficiency as new data formats emerge. Finally, validating the framework's scalability in diverse healthcare environments and data scenarios will be essential to ensure its robustness and practicality in real-world settings, particularly under varying infrastructure conditions. These steps are structured to streamline the framework's real-world applicability and maximize its potential in advancing healthcare innovation.

CONCLUSION

This study demonstrates that a hybrid edge-cloud framework significantly optimizes omics data management, offering a transformative approach that balances efficiency, cost, and privacy factors critical for the advancement of future healthcare applications. The evaluation focused on cytometry data processing efficiency, specifically targeting CLL and B-ALL, and highlighted the advantages of integrating edge computing within this framework. A key finding is that integrating edge computing within this framework significantly reduces data size and upload times. Notably, the ZSTD compression codec was identified as a particularly effective tool, optimizing data-handling capacities in terms of both speed and cost. Although these results are promising, it is important to acknowledge that the study's findings are based on specific data types and settings. The demonstrated benefits highlight the potential of edge computing in omics research, suggesting a valuable direction for future investigations to expand on these findings. Future research will expand this work by developing a machine learning model marketplace for secure, edge-based deployment, enhancing collaboration while maintaining data privacy. Additionally, tailored dynamic compression strategies to evolving genomic and omics data types will be explored to optimize performance. The framework's scalability will be validated across broader genomic and omics applications, such as whole genome sequencing, and tested in diverse clinical environments, including those with limited technologic infrastructure, ensuring practical, real-world applicability.

References

1. Thimbleby H. Technology and the future of healthcare. *J Public Health Res*. 2013;2:jphr.2013.e28.
2. Stoumpos AI, Kitsios F, Talias MA. Digital transformation in healthcare: technology acceptance and its applications. *Int J Environ Res Public Health*. 2023;20:3407.
3. Johnson MO. The shifting landscape of health care: toward a model of health care empowerment. *Am J Public Health*. 2011;101:265–270.
4. Stephens ZD, Lee SY, Faghri F, et al. Big data: astronomical or genomical? *PLoS Biol*. 2015;13:e1002195.
5. Kazemi-Arpanahi H, Shanbehzadeh M, Jelvay S, Bostan H. Developing cardiac electrophysiology ontology:

moving towards data harmonization and integration. *Front Health Inform*. 2020;9:40.

6. Becker M, Worlikar U, Agrawal S, et al. *Scaling Genomics Data Processing With Memory-Driven Computing to Accelerate Computational Biology*. Springer; 2020:328–344.
7. Zou Y, Bui TT, Selvarajoo K. ABioTrans: a biostatistical tool for transcriptomics analysis. *Front Genet*. 2019;10:499.
8. Soneson C. compcodeR—an R package for benchmarking differential expression methods for RNA-seq data. *Bioinformatics*. 2014;30:2517–2518.
9. Velmeshev D, Lally P, Magistri M, Faghihi MA. CANEapp: a user-friendly application for automated next generation transcriptomic data analysis. *BMC Genomics*. 2016;17:49.
10. Habib PT, Alsamman AM, Hamwieh A. BioAnalyzer: bioinformatic software of routinely used tools for analysis of genomic data. *Adv Biosci Biotechnol*. 2019;10:33–41.
11. Nix DA, Di Sera TL, Dalley BK, et al. Next generation tools for genomic data generation, distribution, and visualization. *BMC Bioinformatics*. 2010;11:455.
12. Fisher C. Cloud versus on-premise computing. *Am J Ind Bus Manage*. 2018;08:1991–2006.
13. Mrozek D. A review of Cloud computing technologies for comprehensive microRNA analyses. *Comput Biol Chem*. 2020;88:107365.
14. Hu B, Canon S, Eloe-Fadrosh EA, et al. Challenges in bioinformatics workflows for processing microbiome omics data at scale. *Front Bioinform*. 2022;1:826370.
15. Ali MF, Khan RZ. Distributed computing: an overview. *Int J Adv Netwk Appl*. 2015;7:2630.
16. Reynolds SM, Miller M, Lee P, et al. The ISB Cancer Genomics Cloud: a flexible Cloud-based platform for cancer genomics research. *Cancer Res*. 2017;77:e7–e10.
17. Orechia J, Pathak A, Shi Y, et al. OncDRS: an integrative clinical and genomic data platform for enabling translational research and precision medicine. *Appl Transl Genom*. 2015;6:18–25.
18. Wiewiórka MS, Messina A, Pacholewska A, et al. SparkSeq: fast, scalable and cloud-ready tool for the interactive genomic data analysis with nucleotide precision. *Bioinformatics*. 2014;30:2652–2653.
19. Lesho E, Clifford R, Onmus-Leone F, et al. The challenges of implementing next generation sequencing across a large healthcare system, and the molecular epidemiology and antibiotic susceptibilities of carbapenemase-producing bacteria in the healthcare system of the U.S. Department of Defense. *PLoS One*. 2016;11:e0155770.
20. Krumm N, Hoffman N. Practical estimation of cloud storage costs for clinical genomic data. *Pract Lab Med*. 2020;21:e00168.
21. Leff A, Rayfield JT. Integrator: an architecture for an integrated cloud/on-premise data-service. In: *2015 IEEE International Conference on Web Services*. IEEE; 2015:98–104.
22. Shi W, Cao J, Zhang Q, Li et al. Edge computing: vision and challenges. *IEEE Internet Things J*. 2016;3:637–646.
23. Chethana S, Charan SS, Srihitha V, et al. Comparative analysis of password storage security using double secure hash algorithm. In: 2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon). IEEE; 2022:1–5.
24. Promberger L, Schwemmer R, Fröning H. Characterization of data compression across CPU platforms and accelerators. *Concurr Comput Pract Exp*. 2023;35:e6465.
25. Nguyen QV, Qu Z, Lau CW, et al. Biomedical data analytics and visualisation—a methodological framework. In: *Data Driven Science for Clinically Actionable Knowledge in Diseases*. 1st ed. Chapman and Hall/CRC; 2023:174–196.
26. Smith RG Jr, Windom HL. A solvent extraction technique for determining nanogram per liter concentrations of cadmium, copper, nickel and zinc in sea water. *Anal Chim Acta*. 1980;113:39–46.
27. Belov V, Tatarintsev A, Nikulchev E. Choosing a data storage format in the apache hadoop system based on experimental evaluation using apache spark. *Symmetry*. 2021;13:195.
28. Malta M, Cardoso LO, Bastos FI, Magnanini MMF, Silva CMFPd. STROBE initiative: guidelines on reporting observational studies. *Rev Saude Publica*. 2010;44:559–565.
29. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526:68–74.
30. Weinstein JN, Collisson EA, Mills GB, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet*. 2013;45:1113–1120.