Deep Science Publishing
https://doi.org/10.70593/978-93-49307-76-6_7

DeepScience
Open Access Books

Chapter 7: Cloud computing and scalable artificial intelligence architectures for genomic analytics: Platforms, pipelines, and security

7.1 Introduction

Cloud computing has transformed many industries, including in bioinformatics, on how scientific research and analytics are done today, and those are complementary to disciplines in biology and genetics like genomic analytics. The exponential growth in sequencing technologies has reduced genome sequencing cost by two orders of magnitude compared to 10 years ago, making it increasingly affordable for even small research groups to collect very large datasets. The transformative effect of genomic cloud-computing based solutions for the end-user offers non-expert researchers to safely and efficiently leverage their biological "wet" lab data over massive, distributed, and complex "dry" lab big data through the scalable and cloud-based analysis platforms, services, and tools by "clicking" on the web browser, even on smart devices. Traditional way to analyze data seriously constrains what can be realistically pursued in the time constraints of a single research project, or even several educational and grant cycles. The cost of purchasing the necessary software and hardware infrastructure is well out of reach for small research groups or under-funded academic institutions. Even impressive HPC resources may not be enough to tackle complex questions and perform sophisticated analyses that rival state-of-the-art research, and be under-utilized for the routine daily needs of biologists, in most cases the end-users of genomics data. In some cases, because of bioethical and administrative reasons, it is impossible to send data over the Internet or the requested services can't be found on the market. Prohibitive cost and difficult recruitment of capable bioinformaticians in these regions presents a bottleneck in exploiting these datasets and creates a widening gap in biomedical research to major

research centers in more developed countries. Spark, a major technology used in this work and the most popular computing system for big data processing, allows for a speed-up of factor of 100 times compared to traditional disk-based systems. Thirty-six new software tools and more than 36,000 Genomic Databases were released in the third year. Pre-existing analytics tools will be adapted, addressed, and improved to work within the framework of the Genomic Open-data Architecture (GOA) within the PaNlab integrated hardware-software-stack. The final high-level Object Oriented Cloud Interface (OCI) will be produced that will enable any user to leverage all available services and tools in their own “cloud agnostic” environment securely and efficiently.

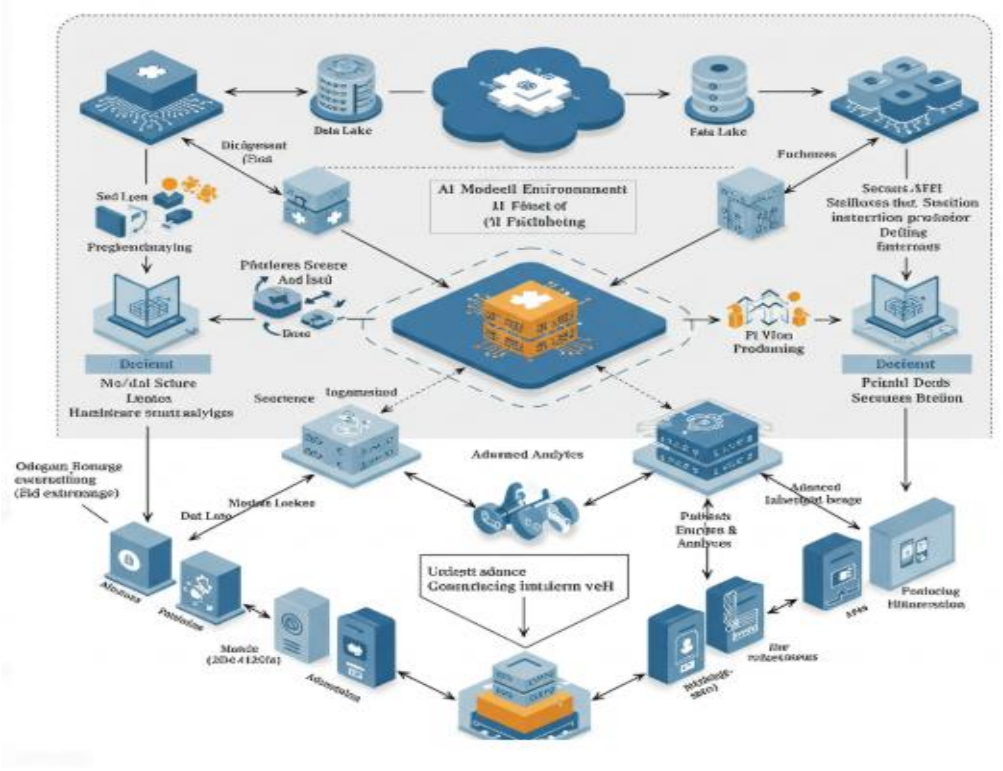

Fig 7.1: Cloud-Native AI Architecture for Scalable and Secure Healthcare Analytics

7.2. Overview of Cloud Computing

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. Cloud computing’s fundamental principles include on-demand self-service, broad network access, resource pooling, rapid elasticity/massive expandability, and measured service/pay-per-use consumption models. As such, cloud computing

represents a convenient and increasingly popular way to expand upon platform capabilities without significant infrastructure investment. A perspective on the cloud computing landscape is provided, including its essential characteristics, deployment models, and service models (Chakilam et al., 2022; Chava, 2023; Challa, 2024).

Taken as a whole, the discussion illustrates a growing appreciation of the cloud among diverse industries and of the economic and operational advantages of cloud-based solutions, in turn driving the adoption of applications primarily for genomic analytics. Genomic cloud computing, broadly defined, is the use of cloud computing technologies to store, process, analyze, manage, and share large, complex genomics and associated clinical datasets using strong computational resources and software applications hosted on distributed platforms. It has the capacity to support the demand for timely and cost-effective research studies, development of new applications, and infrastructure that can offer the broader genomics community access to a wealth of data. A pan-cancer project would ideally allow researchers remote and secure access to raw sequencing data as part of harmonized comprehensive analyses alongside other multi-focused large-scale community projects. To achieve this, one project has addressed a set of legal and ethical issues regarding identifiability for analysis of data for and from large-scale projects in order to maintain the privacy of the individuals and ensure their anonymity. Topics include defining curiosity-driven and hypothesis-driven analyses, handling of controlled-tier data, ensuring secure communication with the research platforms, and drafting of Data Use Certificates.

7.3. AI Architectures in Genomic Analytics

This section develops evolving architectures of artificial intelligence for genomic analytics from a platform and operational perspective and discusses their implications for making genomics data, storage, analysis, and interpretation securely scalable. With the rapid growth of genomic and clinical data, large scalable infrastructures for genome analytics are crucial for diagnostics, drug discovery, and treatment. Today, several software platforms are available for genomic data processing. They are increasingly being used in combination with services available via cloud computing or HPC interfaces for statistical analysis, machine learning, and modeling, enabling the development of intelligent healthcare solutions. This section raises timely issues concerning the platforms, pipelines, and security of analysis results, expands potential challenges and research directions for the further development of scalable AI architectures for genomic data in partnership with existing techniques in cloud computing and federated learning. The outlook includes predicting the growing need for data protection regulations and patient data harmonization requirements among medical centers. It highlights a need for further consideration and development of secure and fair

data markets considering the monetization of genomics data. The section raises themes concerning the important development of scalable architectures for genomics data and the analysis, along with their options to be made securely and fairly scalable. It delves into novel aspects of the subject, like how cloud computing models and common algorithms are transferred to genomics applications and how the results acquired are intended to be protected and shared with the research and innovation community, thus creating a dynamic architecture for the development of advanced intelligent solutions and expanding its beneficial societal impact (Malempati et al., 2022; Nuka et al., 2022; Komaragiri, 2024).

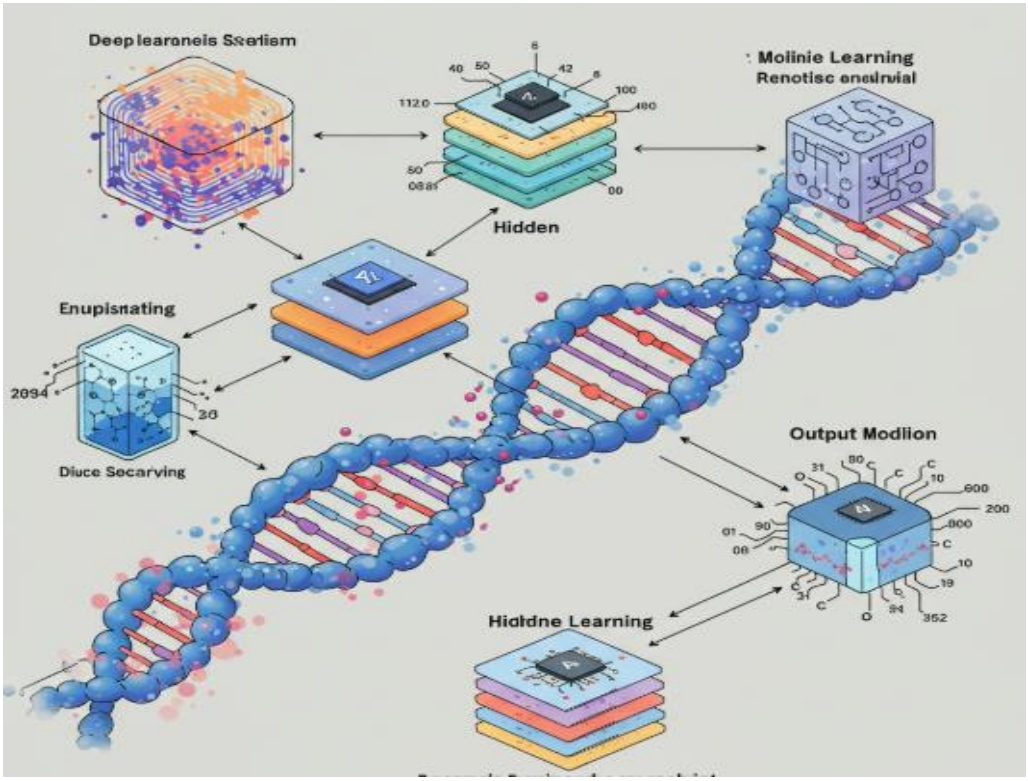

Fig 7.2: AI Architectures in Genomic Analytics

7.3.1. Machine Learning Techniques

This sub sector focuses explicitly on machine learning techniques utilized within genomic analytics, detailing various algorithms and their applications. Both common and advanced topics are explored. Genomic analytics is a burgeoning field of bioinformatics that concerns various aspects of genome sequence interpretation on individual and population scales. While biomedical breakthroughs are being made in genomics including molecular diagnostics, targeted therapies, and personalized

medicine, the gargantuan size and complexity of DNA data bring about unique computational challenges. Here, it focuses on machine learning and scalable artificial intelligence (AI) in dissecting genomic sequences, covering an array of platforms, pipelines, techniques, and applications, leveraging cloud computing resources particularly. It starts by elaborating on standard data formats in genomic analytics, followed by an in-depth discussion of machine learning techniques including regression-based, clustering, and decision trees algorithms. Case studies and example codes are provided, demonstrating how machine learning is used to interpret genomic data and create predictive models. Importantly, consideration is given to feature selection, data preprocessing, and model hyperparameters in ensuring the accuracy and performance of machine learning algorithms within this domain. Commonly used tools and frameworks in scalable AI or machine learning pertaining to genomics are also briefly elucidated. Several challenges of applying machine learning to genomics including overfitting and interpretability are discussed further. Topics range from open-access resources and baseline algorithms to advanced methods and unpublished results that are concrete and informative. It is concluded that machine learning has a profound and transformative role in genomic data analysis and model prediction, which aligns with a recurring vision for personalized medicine and targeted therapies. The groundwork is laid for delving into deep learning and its advanced applications in the genomics that follow.

Understanding the human genome is central to interpreting its biological functions in health and disease. The completion of the Human Genome Project in early 2001 marked an inflection point in scientific discovery, opening the door to high-throughput DNA sequencing technologies that generate valuable data for both basic research and clinical applications. Such data may encompass the entire genome, the exome, or the transcriptome, produced from various omic assays and expressed in diverse biological states. Spotlight initially sheds light on the implications of non-coding DNA and its derived non-coding RNA in regulating cell functions and subsequent diseases. Other research questions on large-scale polymorphic regions and their effects on individual phenotypes, pathway-level interactions and their causality with complex diseases, and population-scale mutations and their dynamics in evolution emerge gradually. The rapid accumulation of genomic data, especially within the big-data context of hundreds of terabytes or petabytes of raw sequences, challenges traditional bioinformatics tools that are computationally intractable or conceptually inadequate.

7.3.2. Deep Learning Applications

Amid rapid advances in biomedical science and genetics, it is reported that the pace of scientific discovery in genomics now significantly exceeds that possible in computing, exacerbating an ever-widening gap between data production and analysis. To help bridge

this gap, this review surveys scalable and secure cloud-based platforms and pipelines optimized for genomic analytics. These cloud-based platforms encompass three types, namely servers equipped with GPU accelerators, scalable multi-server architectures, and serverless computing models. Their hosted pipelines are characterized by interoperable modules, parallel/distributed processing, and assistive security mechanisms. Case studies are presented for each platform, including increasingly complex pipelines for alignment, variant calling, annotation, and scoring. The latter accounts for adverse security implications associated with automated processing and sharing of highly personal data. However, it is strongly argued these risks outweigh societal benefits from the data-driven personalized healthcare revolution. Possible solutions include a framework for secure multiparty computation for secure machine learning. Collectively, scalable AI architectures and pipelines for genomic analytics on the cloud embody an innovative, secure, and ethical roadmap for the realization of technologically and economically viable precision genomics.

7.4. Scalability in Genomic Data Processing

A critical aspect of genomic analytics is scalability in data processing: it must manage the ever-growing volume and complexity of genomic information and its various formats. Genomic data comes from sequencing many thousands of subjects, in many cases at full genomic level, and involves aggregating other types of information (health records, imaging, etc.). It is more complex than other types of Big Data because it is of higher and more diverse dimensionality (for example, a full genomic set contains up to three billion base pairs). There is a critical need to move away from the use of traditional tools and methods for big data in order to realize the full potential of genomic analytics. Progress in DNA sequencing technology is creating datasets growing faster than compute power according to Moore's Law, the other way around to big data development. Regarding these datasets, the biggest interoperability challenge is their diversity, as there are a plethora of specialized data formats and schemas across storage protocols and file sizes. Standardized file-linked formats – like SAM/BAM or VCF – solved it partially for some types of files, but overall most genomic datasets require laborious data wrangling actions due to mismatches in them. Improvements in cloud-based platforms have the potential to solve this diversity issue in genomic interoperability. By populating cloud-based platforms with genomic datasets under standardized formats and schemas, scalable solutions can be built on top. There are still unresolved issues, like existing legal barriers. However, despite them, their entrance might foster further advances in genomic interoperability through widespread possibilities of integrated analytics. Troubling issues rooted in intellectual property of medical data should be mitigated for development in the field and the benefit of science and public health.

A basic survey on the current state of art on the main big data technologies and systems providing solutions to this demand, mostly new hardware architectures and updated programming models, is addressed to provide a high level overview of it. Then, and focusing on scalable architectures, three paradigms providing support to designed architectures meant for big storage, big computing power and genomics crimes is discussed as a domain related to big data on genomics. Four different and popular platforms from the current state of art are subsequently reviewed: GlusterFS, Hadoop, Spark and Cassandra. Pipelines or graphs of tasks with directionality and mapping of the previous architectures are an illustration. Two use cases showing the process of analyzing big data from genomics crimes by running different tasks of the pipeline on those platforms used as archetypes or with an analog architecture to the chosen ones are shown. Results are presented by taking into account the review of the platforms and the examples or case studies above stated. An inter-field analysis is tried.

7.5. Cloud Platforms for Genomic Analytics

Aspects such as scalability, reusability, and adaptability can further motivate the use of cloud computing for scientific research, as opposed to traditional high-performance computing clusters. Although there exist many cloud platforms specifically designed for genomic data analysis, the use of generic cloud providers may be preferred in some cases. Such cloud platforms can offer an extensive range of computational resources and services, enabling users to develop flexible and scalable architectures that can process and analyze the data. Of course, the cloud platform (whether generic or tailored to genomics) must offer tools suited for this high-performance environment, allowing users to rapidly and intuitively analyze together large genomics and omics datasets.

The rise of high-throughput experimental platforms has drastically changed the way scientific research in bioinformatics and biology is carried out. What used to require weeks and even months of labor can now be done in a matter of days and hours. As a consequence, the complete DNA sequences of organisms (their genomes), as well as the DNA variation between them, started accumulating at an unprecedented rate. The scientific community was not ready for the data deluge on the one hand, and also unprepared for the potential of the new data streams on the other. Aspects such as data storage, transfer, and processing started becoming bottlenecks. It was clear that traditional computing biology models were no longer adequate for this kind of data. Also, the fact that this situation is a consequence of the speed of the experimental platforms highlights the fact that this is a technological, and not a scientific bottleneck. Indeed, with the right tools, only a few advanced training is required for wet lab biologists to also be able to operate and analyze the new data generation platforms. As a consequence, many initiatives in this decade have put forth biosciences and

bioinformatics strategies to process this data. But it became clear very fast that the phenomenon was well beyond the reach of even those collectively envisaged strategies. In this scenario, the rise of cloud computing platforms also oriented to the providing of resources for the scientific community can cover this gap. For one, the needed infrastructure can be easily acquired: one can launch new storage drives, cluster nodes with the data and bioinformatics software stack needed; secondly, as it will be seen below, cloud resources are matched very closely to the big data/parallel processing paradigm in bioinformatics. For those reasons it was endeavored to explore a platform that would enable the deployment of state of the art genomics and omics analytics pipelines in a cloud provider.

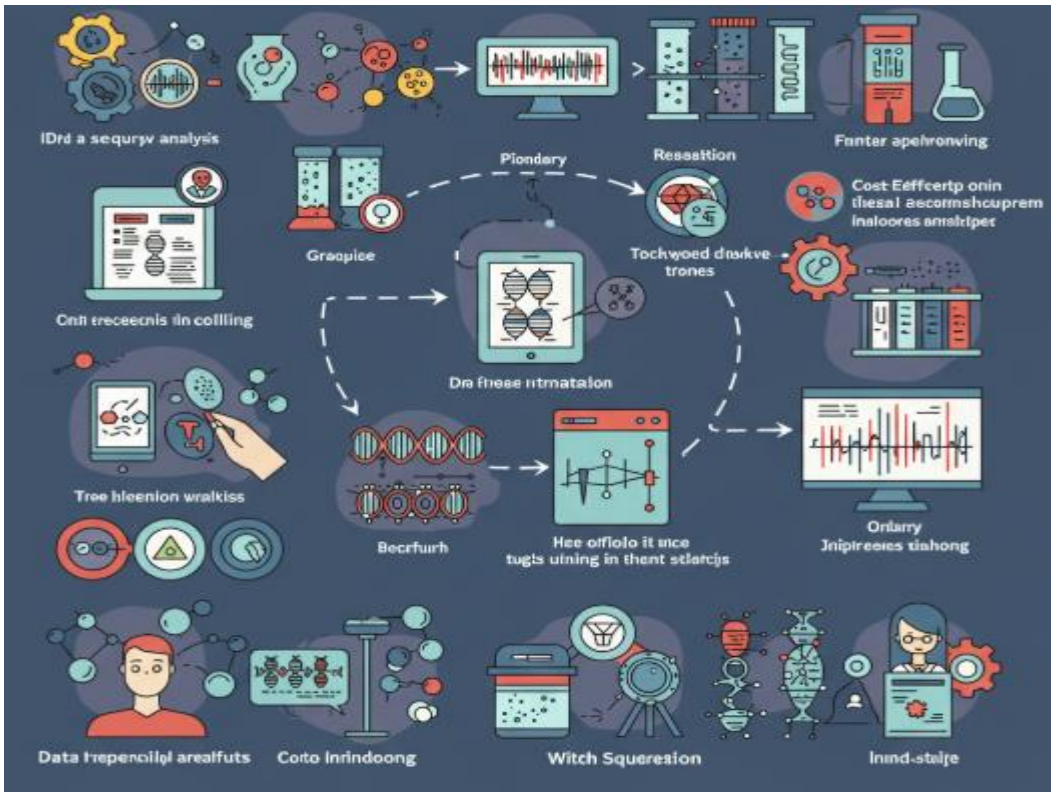

Fig : Cost-effective and accurate genomics analysis

7.5.1. Amazon Web Services (AWS)

Genomic cloud computing is an emerging paradigm that involves the utilization of cloud computing technologies for the acquisition, storage, and analysis of large-scale genomics data. The major providers of cloud computing include . All of these platforms offer a wide variety of services that are useful for both first (primary) and second (secondary)

analysis of genomic data. Here, cloud computing is introduced, the bioinformatic tools and services that are supported by these services are highlighted, and there is an exploration as to how the use of this technology is revolutionizing the field of genomics. Cloud computing services are complemented by the bioinformatic tools and pipelines that are available to researchers for the FASTQ and VCF (and related) file formats across different genomic tasks. are provided in extensible architecture that allows developers to integrate new bioinformatic servers and services. This provides an avenue where genomic data and bioinformatic software applications can be better integrated to harness the enormous opportunity of big genomics data analytics while reducing the workload & expertise thresholds otherwise required from a resource-poor (scientific) end user. Many recent studies have demonstrated the great utility of cloud computing for big data intensive genomics analysis since high performance computing and storage capabilities directly improve the performance of these applications, removing existing limitations.

7.5.2. Google Cloud Platform (GCP)

Rapid advancements in high-throughput sequencing technologies and genomics offer a comprehensive understanding of the structure, function, and evolution of genomes. However, the resulting large-scale genomic datasets, characterized by terabytes to petabytes in size and millions to billions of records, have surpassed the capability of traditional computing resources when conducting complex bioinformatics operations. While several compute and storage solutions may help scientists face these challenges, cloud computing technology is highlighted for its adaptiveness to growing data scales, easy accessibility, and favorable cost-efficiency.

Bioinformatics researchers can launch and manage their pre-built environments, write pipelines, and execute their workflows directly on these remote machines. It is anticipated that these projects will drive users to consider running their pipelines on cloud resources more frequently in the future. Here there are a discussion of the more technical aspects of cloud computing and its applicability to genomic and bioinformatics workloads. There are also guidelines to help set up cloud computing workflows as easily and effectively as possible.

7.5.3. Microsoft Azure

The methodologies to report on descriptive analytics do not advance as rapidly as the improvements seen in algorithmic techniques. Yet, several recent innovations have been adopted that can enhance the best practices currently in place. The unannotated lateral swap, above; other tools, including multi-faceted visualizations, safety measurements and reporting recommendations; and the proposal for the open publication of machine-

learning specifications (including their chosen settings for random number generators) could contribute to raising awareness of the technique and promoting its detailed reporting. Each of these improvements is intended to lower the barrier to entry for the non-specialist user of ML technologies and to promote basic confidence in the assignment of feature importances. In addition to other technical advances for conservative and global sensitivity analysis of trained machine-learning models, a new, pragmatic form of analysis called the unannotated lateral swap is described that may be widely applied to the machine-learning systems now routinely used in genomics and bioinformatics. While such a sensitivity analysis is not able to return a summary metric by which to rank genome–phenome associations, it can significantly augment the findings of any analysis that does not contain such a sensitivity analysis. To this end, best practice policies are proposed for the reporting of machine-learning analyses that are actionable by both authors and peer reviewers.

7.6. Conclusion

The transformative impact of cloud computing and scalable AI architectures on genomic analytics research and industry is discussed in detail. Scalable platforms for analytics of integrated genomes and health records are presented, benchmarked using population-scale resources, to the characterization of human complex diseases, which are now being recognized as perturbed dynamic biological networks. Scalable machine learning pipelines for analyzing genomes, epigenomes, microbiomes, and transcriptomes—frequent perturbing interactions—across hundreds of thousands of samples are presented, benchmarked using relevant resources, and applied to identify potential causal mechanisms underlying genome-wide association signals. A method is introduced that only requires access to summary statistics, enabling "genome-wide-transcriptome" and "epigenome-wide-transcriptome" atlas of perturbation interactions in humans at near-zero marginal cost. Finally, legal and ethical points to consider in leveraging cloud computing for big data genomics are discussed.

In 2003, the first human genome was sequenced at a cost of US$0.5 billion. Recently, a transformative milestone was reached when a new technology was announced, which could generate a human genome at a sequencing cost of US$1000 with a single flow cell. By the time when the genotyping is completed, the sequencing cost may further decrease. In conjunction with the availability of scalable and powerful computing frameworks, new analysis tools, and efficient methods, scalable in-depth artificial intelligence (AI) architectures. At the transformative moment, in the era of ultra-low cost and massive-scale human genetic sequencing, a research agenda, albeit speculative, is presented. This essay is broadly segregated into five parts.

The first section is Cloud Computing and Scalable AI Architectures for Genomic Analytics, which describes the transformative impact of cloud computing and AI architectures on genomic analytics research and industry. Given a multitude of topics, they are carefully and deliberately chosen to represent different facets of the landscape. Two examples are presented to illustrate the power of large-scale machine learning analytics and platforms in (1) characterizing complex diseases as perturbed dynamic biological networks and (2) identifying potential causal mechanisms underlying genome-wide association signal perturbations.

7.6.1. Future Trends

The worldwide need for interdisciplinary research in cloud computing and scalable AI architectures for genomic analytics is generally framed and motivated. The genomic use case is contextualized, in concert with an illustrative overview of current scientific studies in the field of population and cancer genomics. The main part of this entry then builds on the relevant technological and scientific developments, delving into four aspects, each accompanied by its relevant scientific contributions. Future trends in cloud computing and scalable AI architectures for genomic analytics are discussed, followed by concluding remarks. The growing body of research in the field is monumental. Cloud computing has emerged as a cost-effective and scalable technology in genomic data analysis and sharing in response to the exponentially increasing growth of genomic data. Innovations in scalable artificial intelligence architectures are analyzed, with a focus on fast and accurate analytics. Big data, in terms of both volume and variety, and the emergence of social media, Internet of Things, and data-pervasive science add to the challenges. Practical aspects relate to large-scale integration and interoperability between various genomic platforms. Population and patient genomics are foreseen in daily tasks in the near future, as the plummet of sequencing costs continues. The anticipated boom coincides with recent breakthroughs in cancer genomics, such as single-cell sequencing technology and open-access datasets. As current computational biology research rarely covers such scale and complexity, sharing resources and expertise to build a state-of-the-art cloud-based pipeline and platform is vital. Leading role is suggested for collaboration between academia and the industry on software and hardware innovation to address emerging computing and data-management challenges. Quantum revolution may further boost the fast growing and high-dimensional genomic data exploration; however, it dovetails with concerns over data privacy and security. Legal and ethical implications of genomic cloud computing in terms of data control and protection are briefly discussed. The joint development of a genomic cloud data platform is advocated. Dedicated efforts in research and development are acknowledged in the final note to fruitfully embrace this data-intensive bioinformatics era.

References

Chava, K. (2023). Revolutionizing Patient Outcomes with AI-Powered Generative Models: A New Paradigm in Specialty Pharmacy and Automated Distribution Systems. Journal for ReAttach Therapy and Developmental Diversities. Green Publication. https://doi. org/10.53555/jrtdd. v6i10s (2), 3448.

Komaragiri, V. B. (2024). Generative AI-Powered Service Operating Systems: A Comprehensive Study of Neural Network Applications for Intelligent Data Management and Service Optimization. Journal of Computational Analysis and Applications (JoCAAA), 33(08), 1841-1856.

Chakilam, C. (2022). Integrating Generative AI Models And Machine Learning Algorithms For Optimizing Clinical Trial Matching And Accessibility In Precision Medicine. Migration Letters, 19, 1918-1933. Malempati, M. (2022). Machine Learning and Generative Neural Networks in Adaptive Risk Management: Pioneering Secure Financial Frameworks. Kurdish Studies. Green Publication. https://doi. org/10.53555/ks. v10i2, 3718.

Challa, K. (2024). Neural Networks in Inclusive Financial Systems: Generative AI for Bridging the Gap Between Technology and Socioeconomic Equity. MSW Management Journal, 34(2), 749-763.

Nuka, S. T. (2022). The Role of AI Driven Clinical Research in Medical Device Development: A Data Driven Approach to Regulatory Compliance and Quality Assurance. Global Journal of Medical Case Reports, 2(1), 1275.