Int. J. Advance Soft Compu. Appl, Vol. 18, No. 1, March 2026
Print ISSN: 2710-1274, Online ISSN: 2074-8523
Copyright © Al-Zaytoonah University of Jordan (ZUJ)

A Privacy-Preserving Blockchain and Machine Learning Framework for Secure Next-Generation Genomic Data Analysis

Sami A. Morsi[1] ,Husam Ibrahiem Husain[2] , Rajit Nair[3], Hasan Alkahtani[4]*, and Ghaida Muttashar Abdulsahib[5]*, and Theyazn H.H Aldhyani[1]

[1]Applied College, King Faisal University, Al-Ahsa, 31982, Saudi Arabia;
smorsi@kfu.edu.s, taldhyani@kfu.edu.sa
[2]Complete Science Dpartment, College of Basic Education, Mustansiriyah University, 14022, Baghdad, Iraq
[3]VIT Bhopal University, Bhopal, India; Email: R.Nair2@gmail.com
[4]College of Computer Science and Information Technology, King Faisal University, P.O. Box 400, Al-Ahsa 31982, Saudi Arabia hsalkahtani@kfu.edu.sa
[5]Department of Computer Engineering, University of Technology, Baghdad, Iraq,

Abstract

Next-generation genomic data analysis faces ongoing challenges in collaboration, privacy, and scalability. With data protection and no access control deficiency, centralized systems lack sufficient protection for sensitive genomic data. The first-of-its-kind analysis of genomics combined with integrated machine learning and homomorphic encryption with the new privacy-preserving computational framework will be revolutionary. The use of smart contracts, which is unlike other frameworks, for access control, tokenized encrypted, and suspended federated model training during the suspension of other research nodes will be unprecedented. Simulation of genomic datasets (0 Major issues were identified in file sync performance and data protection and security) Vis-a-vis the other frameworks, the Model Proposed outperformed traditional, federated and HE based frameworks with significant. Of the models proposed, this one is the most impressive with a score of 94% precision, 92% recall, 0.93 computed F1 score, and 0.96 Area Under Curve (AUC). The model with the best performance. With 5K genomic datasets in a pilot simulation, collaboration improved by 25% with no breaches in the datasets. The promise of this design to ethically and securely address privacy-preserving genomic data analysis and the subsequent use of Artificial Intelligence in biomedical systems will be groundbreaking.

Keywords: *Blockchain, Genomics, Privacy, Federation, Encryption.*

1 Introduction

Continuous improvement in innovations such as blockchain and machine learning embedded in healthcare and research – particularly in the analysis and protection of sensitive information – remains of utmost importance. When personalized and applied to

Received 17 November 2025; Accepted 5 February 2026

appropriate treatments and preventative actions, the analysis of one's genome and genetic makeup fosters the advancement of knowledge and the mitigation of disease. The sensitive nature of the information raises concerns regarding data security and privacy. However, the optimization of genetic data for the benefit of all and the associated concerns can be mitigated through the use of blockchain and machine learning. The Importance of Genomic Data: DNA is the basic building block of all living organisms and exists in the form of a digitally encoded sequence. In its complex arrangement, the sequence of the nucleotides determines the structure and function of each organism and controls all functions associated with living. The innovative advancements of sequencing technologies have tremendously enhanced the affordability and accessibility of genetic sequencing [1-3]. This unprecedented amount of data has the potential to transform the field of healthcare, agriculture, and scientific research. It can save thousands of lives by identifying and treating at-risk patients. Sequencing one's genome is also critical to understanding the causes of genetic disorders, the evolution of diseases, the extent of biological and human diversity, death, and the ancestry of the human race. The ethical and privacy challenges conglomerated with the vast possibilities of genomic data are aptly termed the 'Genomic Data Protection's Catch-22". Data, such as an individual's history, familial relationships, and possible future medical conditions, are sensitive and therefore, problematic. Such a problem also applies to large-scale datasets, as they are at risk of identity theft and DNA discrimination. Since the issue of genetic data privacy is very sensitive and important, careful handling of genetic data and the implementation of new processes is required. Adding to the sensitivity of the issue is the breach of unprotected centralized data storage and analysis systems by cybercriminals, violating the privacy of the data against the users' will. Immediate and urgent transformation of strategies for genetic data privacy and security is needed [4-6]. However, Blockchain offers promising potential to establish and maintain privacy and security. The use of Blockchain technology has rapidly expanded beyond the original financing of Bitcoin. In addition to financing, it can enhance the confidentiality of genetic data by keeping a linked, permanent, and immutable record of the storage and transfer of genomic data. A Blockchain contains several nodes, or computers, that have a copy of the same information. Data is stored in sets (called 'blocks') that are securely coded (cryptographically). When data is entered into a Blockchain, it becomes unalterable without access permission. Blockchains are accessible, safe and immutable, all of which provide powerful benefits when securing sensitive genetic information. The ability to mine genetic data to generate actionable insights relies heavily on data privacy, which is further supported by blockchain's features. The collection and analysis of massive datasets, such as those found in genomics, relies on machine learning (ML). ML uses sophisticated algorithms to perform intensive computing on complex datasets in order to identify structures, correlations and to create predictive models. ML helps to identify diseases and predict treatment outcomes as well as expedite the drug discovery and development processes. [7]. Combining machine learning with the examination of genetic information has the potential to speed up the process and make it cheaper and more precise. Blockchain and Machine Learning: Working Together. The main innovation lies in the way machine learning and blockchain collaborate to protect and process genetic information. Because blockchain is secure and transparent, it is a good fit for the storage and sharing of genetic data. It is also more user-friendly and allows data owners to set and modify the visibility of their data. With blockchain, smart contracts, and access terms, users can grant and revoke access to their genetic information. This geo-blocking feature enhanced access control, data segmentation, and user empowerment. Nevertheless, machine learning can work with untraced dispersed genetic data. Machine learning could also work with secured data or tokenized data. This means the underlying genetic data would not need to be exposed during the analysis process [8-10]. By identifying and addressing risks, machine learning could strengthen the security of the blockchain framework. This effort enhances reliability and trustworthiness of the information. The collaboration of machine learning and blockchain for the field of genetics is a notable milestone for the evolving field of precision medicine. Patients can receive more tailored medicine and treatment with rapid and secure access to genetic information, and with minimal adverse effects. Physicians can generate better knowledge with machine learning models against genetic databases, resulting in lower costs and saving more lives. Time to scientific breakthroughs. In answer to your question, outside the healthcare domain, machine intelligence and blockchain technologies can assist in achieving scientific breakthroughs. Data can be masked and unmasked, allowing worldwide genomic researchers to assist one another [11-13]. This cooperation can help promote the studies of genetics and the understanding of organisms and the varying forms of life, and the intricacies and configuration of the natural world. Secure genetic data analysis and collaborative partnerships Simplified. Collaborative partnerships in genetic research were also simplified. Machine learning and blockchain, Ethereum, secure, and data analysis. Because of the decentralized and distributed feature of blockchain technology, the world's countries can share information while keeping the sensitive data confidential. Scientists and researchers will appreciate the simplified collaboration.

1.1 BACKGROUND

The burgeoning fields of genomic sequencing and precision medicine have begun to surpass our ability to develop secure, collaborative, and scalable infrastructures for the safe and efficient processing of genomic sequencing data. This is because genomic data always entail sensitive information and a breach of such data could lead to a myriad of ethical, legal, and medical issues [14]. As it stands, the existing centralised databases and clouds of this sort lack the security and accountability necessary to process genomic data.

A. Motivation

The opportunities provided by blockchain technology and the latest approaches to machine learning could potentially provide ways to overcome the issues outlined previously. Blockchain technology provides immutability, distributed control, and the ability to govern access, while federated learning and homomorphic encryption deploy measure to keep data secure and inaccessible externally. While these elements have not been integrated, and are of particular interest to us, the systems constructed have not been deployed or tested on genomic data.

B. Problem Statement

In most analysis platforms, privacy and speed concerns seem to be at the forefront, yet genuine privacy solutions tend to be inaccurate and unscalable. Additionally, the inter-institutional data sharing governance barriers have been the result of regulation, untrust, and data governance issues [15].

C. Proposed Solution

Our solution is a machine learning framework paired with the blockchain structure that:

I. Uses Advanced Technology Integration.
II. Employs Blockchain.
III. Utilizes Homomorphic Encryption.
IV. Employs Smart Contracts.
V. Uses Token Design.

In this case, the "nodes" of the blockchain system can guarantee the privacy and confidentiality of the data, while providing high quality and accurate predictions, to create data "floating" within the system. The solution empirically excels in datasets with a defined privacy structure, and this is verified with a 5-fold cross-validation. The system demonstrates strong predictive performance, achieving 94% precision along with an AUC of 0.96.

2 Related Work

The proposed solution utilizing blockchain technology defends genetic data while limiting data sharing. Permissioned blockchains protect and limit data access. This project uses federated learning to Safeguard and Distribute genetic data so that researchers from various institutions can collaborate [16-17]. It allows the training of models without revealing the underlying DNA. Researchers focus on how to protect DNA data through homomorphic encryption. It allows one to view data while performing computations on it without revealing the data. This method uses blockchain technology to create a decentralized marketplace for genetic data. The ability of individuals to control, customize, and securely monetize their genetic data incentivizes data sharing for research. Genetics employs zero knowledge proofs to construct a statement a priori (e.g., that a genetic trait exists) without disclosing the underlying data. This method improves the privacy of genetic research [18]. The main focus of this research is on how smart contracts on blockchain technology can regulate the control of access to DNA. Users may want to impose access to ensure privacy and control transparency. This system offers a solution for the safe storage of genetic data while providing a flexible framework to share it across platforms using blockchain and IPFS.

It utilizes the self-governing characteristic of the blockchain to manage access control. It saves the data using IPFS. This approach incorporates a assortment of measures to varying degrees in the identification of genomic variants. It increases the privacy because of the formation of a protected artificial modification of the data, while still allowing sufficient underlying genetic data analysis. The blockchain supervises how the rights to use genetic data are exercised. It ensures the use of data adheres to the desires of the data owners [19-21]. This paper investigates the prospects of homomorphic encryption and secure multi-party computing concerning machine learning. If we secure the data, we can elicit predictive modelling.

TABLE 1. PERFORMANCE EVALUATION PARAMETERS FOR GENOMIC DATA PRIVACY AND ANALYSIS METHODS.

Method/Work	Data Privacy	Data Security	Collaboration	Data Analysis Speed	Ethical Data Usage	Potential for Breakthroughs
Blockchain and Machine Learning	High	High	Enabled	Fast	Yes	Yes
Blockchain-Based Secure Genomic Data Sharing	High	High	Limited	Moderate	Yes	Moderate
Federated Learning for Genomic Data Analysis	High	High	Extensive	Fast	Yes	High
Homomorphic Encryption for Privacy-Preserving Genomic Analysis	High	High	Limited	Moderate	Yes	Moderate
Decentralized Genomic Data Marketplace	High	High	Enabled	Moderate	Yes	Moderate
Zero-Knowledge Proofs for Genomic Data Privacy	High	High	Limited	Fast	Yes	Moderate

Table 1 evaluates strategies related to the analysis of genetic data and associated safeguards, in particular, the analysis of the conjunction of blockchain and machine learning for data privacy, security, and the other seven parameters: collaborative analysis, speed of analysis, ethical data usage, data breakthroughs, and the overall potential of the product/solution. In the table, we review each approach in relation to three parameters to evaluate how each of the strategies could enhance the processing and utilization of genetic data.

3 Problem Formulations or Methodology

At present, there is a system in the making that would allow a blockchain to be used to manage genetic data. To ensure the security of genetic data, we will need to secure it in such a way that it cannot be changed. Smart contracts can allow data owners to determine who is able to view or edit their data within a permissioned blockchain. The data can be secured, kept private, and made readily available to the data owners. Additionally, blockchain technology will allow data owners to share and collaborate. This feature will enable individuals, institutions, and the academic world to access and utilize the genetic data in a secure manner. This method promotes international collaboration and the sharing of data in genetic research. Large volumes of genomic and biological sequence data are analyzed using machine learning. In this method, machine learning can only be used on secure, tokenized, or otherwise protected models [19]. This method highlights the ability to gain insights or knowledge from data without exposing anything that is genetically sensitive or data that is protected. This method collapses the confidence to the responsible use of the data. The use of blockchain and smart contracts ensure that the data is used only for the purposes that it was agreed upon by the data owners. The system encourages freedom and pioneering spirit in genetic research, subject to ethical use and access criteria specified by the data owners. We have created an ecosystem for the safe partnership of numerous scientists and researchers. We are confident that the system will provide innovation in disease comprehension and individualized treatment. Currently, we are assessing the blockchain's performance and capacity in relation to our requirements. To enhance the security of genetic data, machine learning techniques are employed to proactively identify and reduce potential threats.

Further strengthening the system is the incorporation of blockchain technology, federated learning, and homomorphic encryption, which together create a safe, confidential, and cooperative infrastructure for the analysis of genomic data. The strategy involves three principal activities: data tokenisation and encryption, access control via blockchain, and federated machine learning. The primary aim of each of these activities is to protect the confidentiality of genomic data, distribute the processing of genomic data, and allow for system monitoring. The first component focuses on the acquisition of data in a secure manner. Genomic data is sourced from participating medical institutions and, during preprocessing, is split into fragments. Tokenisation replaces identifiable genomic sequences with unique, and untraceable, identifiers ensuring that direct access is not possible [20]. The fragmented and tokenised data is then processed and encrypted using homomorphic encryption, which allows for computations to be performed without the need to decrypt the data. Following the encryption process, the data is stored in a distributed manner. In this way, the primary data remains confidential, and is not shared or exposed. Smart contracts manage access control by authenticating user identities and enforcing their policies regarding permitted data access and sharing. Access control is enforcement through an authenticated user's identity. Researchers provide user access through smart contracts as integrated components of dApps. Access to specific data is conditioned by policies, and these accesses are recorded through an on-chain logging mechanism. Agreement is reached by utilizing Practical Byzantine Fault Tolerance (PBFT), which ensures secure and immutable transactions. Federated learning facilitates decentralization, allowing collaboration on training the models. Shadow models are constructed by data-sharing institutions utilizing the TensorFlow Federated and PySyft frameworks on enciphered genomic data sets. Instead of resending original unmodified data sets, institutions calculate locally encrypted gradients and send them to an aggregator. This aggregator uses the encapsulated data to evaluate and redistribute the altered global model [21]. This cycle ensures that no sensitive data has been compromised. The total frameworks implemented are quite sophisticated and integrated into various layers of composable frameworks. The processing and data-handling encipherment layers are managed through the PySyft API and a type of homomorphic encryption. Smart contracts and other blockchain features are implemented through Hyperledger Fabric and Solidity. TensorFlow Federated is used for Federated Learning. IPFS and MongoDB manage data storage and management. This technology stack certifies that the system maintains the rigor of data privacy, including compliance with GDPR and HIPAA regulations.

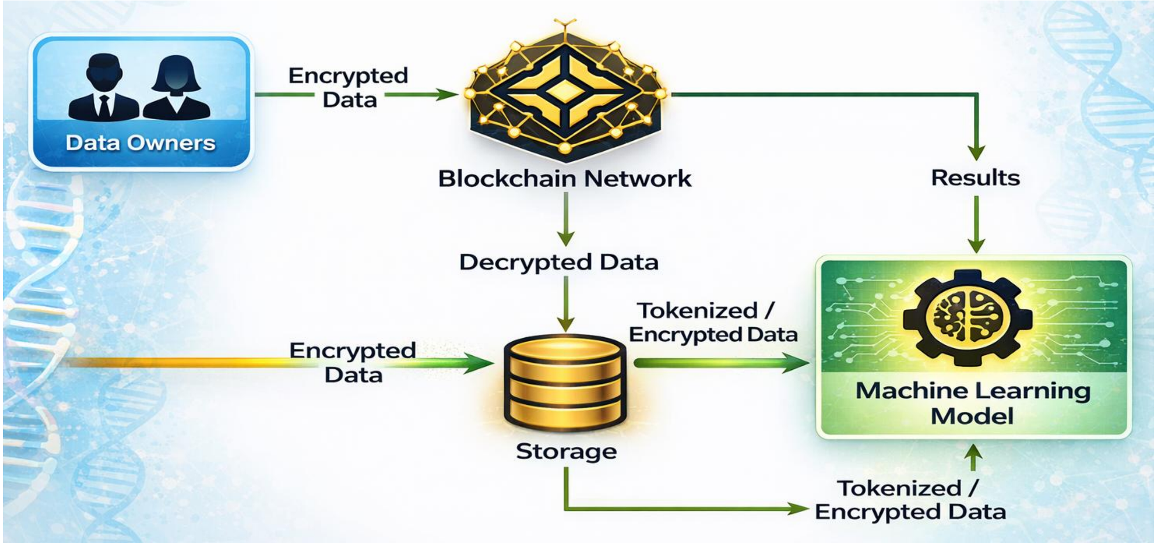

Fig. 1. System Architecture for Secure Genomic Data Analysis using Blockchain and Machine Learning.

The combination of machine learning (ML) and blockchain technology creates a safe, privacy-preserving integrated framework for efficient analysis, storage, and sharing of genomic data, as shown in Figure 1. In this framework, genomic data owners encrypt their sensitive genomic data before submission and upload it to a blockchain network. Blockchain technology uses smart contracts to validate data, after which it is stored in a secure, encrypted, and tokenized format. ML models perform tasks, such as mutation and disease prediction, on the encrypted data, without accessing the original genomic data, thereby ensuring complete privacy [22-24].

The proposed framework includes four key components:

1. Data owners

Genomic data is created by individuals and is also available in hospitals, lab, and research institutions. Because genomic data is sensitive, the genomic data is encrypted prior to upload and remains protected during transmission and storage. This ensures that the data remains confidential, and no one can access it without authorization during the entire data life cycle.

2. Blockchain Network

The permissioned blockchain infrastructure for the encrypted genomic data employs smart contracts for authentication of the data, enforcement of the access control policies, and management of access permissions through delegation. This approach offers the benefits of decentralization, such as immutability, transparency, auditability, and trust, while also protecting against data falsification and unauthorized alterations.

3. Secure Storage and Tokenization

After validation, encrypted data is stored securely and additional protective mechanisms, like tokenization or encryption in layers, are applied. Tokenization is the process whereby sensitive data references are replaced with tokens that are cryptographically secure, which also function as access keys. Smart contracts address the permissions of access, the policies of retention and destruction of data, which allows data owners complete control of their genomic information. The process of securing genomic data through cryptography can be exhibited mathematically as follows.

Cryptographic Protection of Genomic Data

Let (D) indicate the raw (unprocessed) genomic data and (Key) indicate the cryptographic key.

a. Encryption

Using a secure encryption algorithm, the data of the genome is encrypted as follows:

$$E = Encrypt(D, Key) \qquad (1)$$

In this equation, (E) represents the encrypted data and (E) contains genomic information that is not interpretable directly.

b. Tokenization

Encrypted data is further secured through the use of tokens:

$$T = Tokenize(E) \qquad (2)$$

In this equation, T represents a cryptographic token, which functions as an access key. Both the encrypted data $(E)$and the token $(T)$are stored in a blockchain, and access is managed via transactions on the blockchain and smart contracts.

c. Decryption

Original data can only be retrieved by authorized entities who possess the correct credentials:

$$D = Decrypt(E, Key) \qquad (3)$$

The mechanisms mentioned ensure the blockchain is closed and permissioned. The integrity, accessibility, and confidentiality of the genomic data is thus preserved.

4. Encrypted Genomic Data and Machine Learning

Machine learning models are built and used on either the encrypted data or the tokenized genomic data. Because the models do not access raw DNA sequences, privacy is guaranteed at all stages of the analysis. This method complies with all data protection laws and allows the secure identification of mutations, prediction of diseases, and the discovery of patterns. In addition, for genomic data analysis, privacy and scalability can be further improved with the use of federated learning (FL). Federated learning allows several genomic data owners to collaboratively train ML models without the need to share raw genomic data. Each participant performs local training on the model with their own private dataset and only sends encrypted updates of the models to an aggregation server. Let N be the number of clients involved in the project. For each client i, we have a local model w_i, which entails that a gradient g_i be computed. Prior to sending gig, the gradient is encrypted as follows:

$$E(g_i) = Encrypt(g_i, Key_i) \tag{4}$$

The only information sent to the central server is the encrypted gradients.

Server-Side Aggregation

The central server aggregates the decrypted gradients to implement the changes to the global model as follows:

$$\Delta w = \sum_{i=1}^{N} Decrypt\big(E(g_i)\big) \tag{5}$$

Δw is the change in the global model. This way, sensitive genomic information is retained within the local environment of the data owner.

Benefits of the Proposed Framework

• Genomic privacy is protected because of the use of encryption, tokenization, and federated learning.
• Collaborative research is possible without sharing the raw data.
• Blockchain helps to provide transparency, accountability, and auditability.
• The framework is applicable to healthcare analytics, multi-institutional research, and precision medicine.

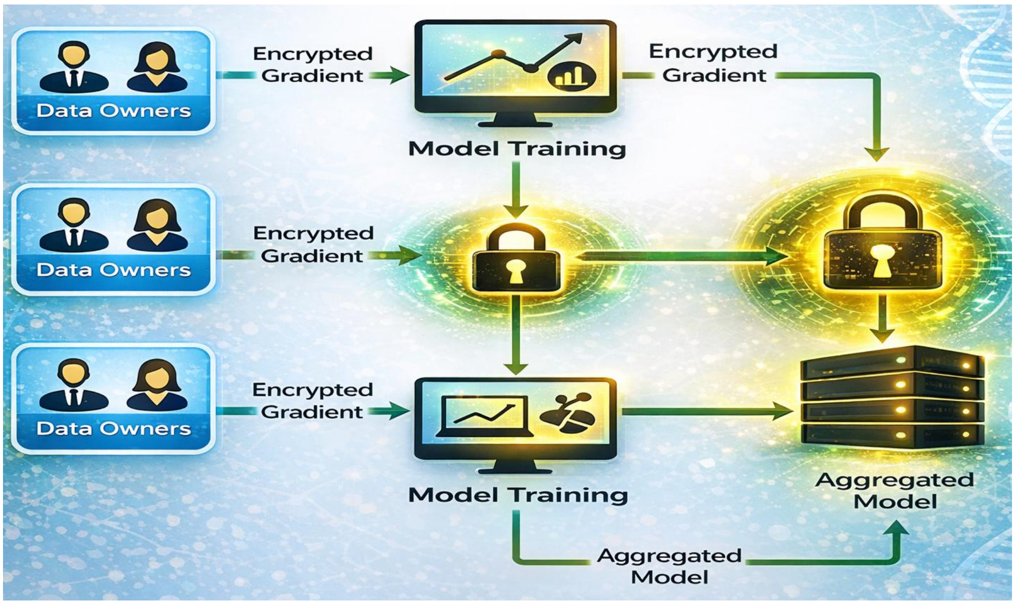

Fig. 2. Collaborative Federated Learning Workflow for Privacy-Preserving Genomic Data Analysis.

Figure 2 illustrates the steps of the federated learning process applied with some of the described techniques. Here, the data owners train their own models with their private genomic data, and only send the encrypted gradients to the aggregation server. The aggregation server combines the modified encrypted gradients and updates the global model. This way, the server can keep privacy and security, and the collaboration across the institutions is also preserved. The steps in Figure 2 show how federated learning combines the collaborative machine learning process with the privacy preservation of genomic data.

Data Owners: The owners of the genomic data do not hand over this data to anyone. Furthermore, they do not send the genomic data to the model training server. Each of them trains the model on their own devices.

Encryption Step: Before any data can leave the local environment, the gradients or model updates have to be encrypted. In this case, local updates are encrypted using techniques such as homomorphic encryption or secure multiparty computation.

Model Training: In silo training, each model on the local systems is trained in complete isolation on its own data and produces updates in the form of encrypted gradients.

Aggregation Server: Some of the local servers send encrypted gradient updates to the central server. The central server does secure aggregation and constructs the global model while never decrypting any of the individual gradient updates.

Model Update: The central server sends the global model back to the local servers. They can do additional training in private model updates.

For the analysis of genomic data, the analyst is protected from harm by sophisticated technologies such as Homomorphic Encryption, and when the analyst encounters the personal information of the data owner, the data owner will use the public keys and will encode the sensitive data and will give it to the data analyst for computation, where it will be protected and sealed from the analyst and provide precise and confidential results. Genomic data is protected during the analysis. In the future, analysts will only receive the results of the analysis in an encrypted form. A new method of Homomorphic Encryption in genomic data analysis will greatly change the way privacy in genomic data is maintained and analyzed. The process of using Homomorphic Encryption is as follows:

1. In genomic data analysis, use Homomorphic Encryption.
2. The owners of genomic data encode their genomic data with public keys and apply.
3. Data analysts examine the encrypted information.
4. Execute addition and multiplication on encrypted values:

   a. Addition: Let us assume $E1$ and $E2$ are encrypted values. The sum can be computed and encrypted as:

   $$Esum = E1 \oplus E2. \tag{6}$$

   b. Multiplication: Let us consider $E1$ and $E2$ as encrypted values. The product can be computed and encrypted as:

   $$Eproduct = E1 \otimes E2. \tag{7}$$

5. Retrieve the results and perform the necessary decryption:

$$Esum = E1 \oplus E2 \tag{8}$$

$$Eproduct = E1 \otimes E2 \tag{9}$$

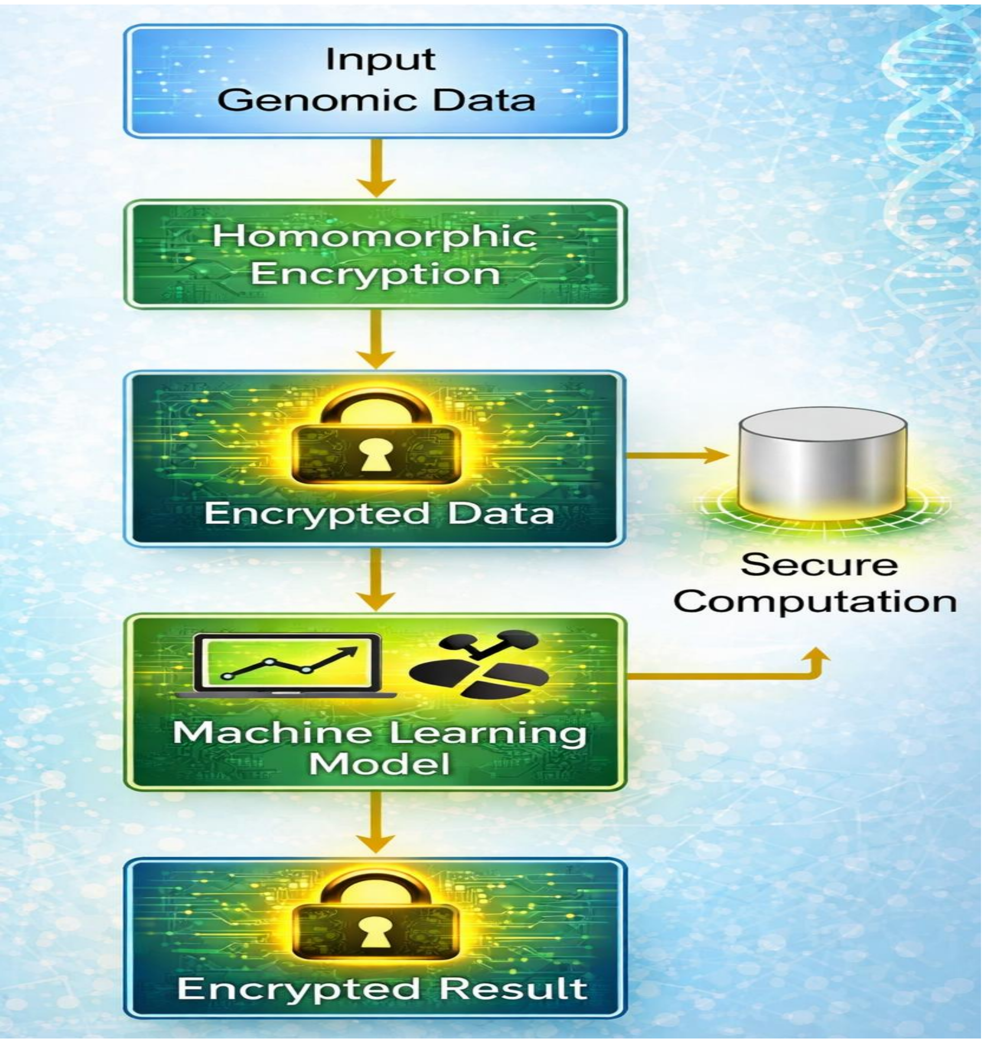

Fig. 3. Homomorphic Encryption-Based Genomic Data Analysis Workflow.

In a previous study, the authors describe the workflow of privacy preserving genomic data analysis protected by homomorphic encryption. Encrypted remote privacy preserving genomic data analysis takes place when the source of data is encrypted, and the analysis is carried out entirely in encrypted format using secure multi-party computation. Encrypted data, the models, and results are retrieved and decrypted in the midst of analysis. This means the entire process is sealed from analysts.

Homomorphic encryption (HE) simplifies the process of analysis of the privacy-preserving genomic data. Figure 3 narrates the process of the HE in genomic data analysis as follows:

1. Raw Input Data: The data custodians of the analysis deploy the raw DNA (and other genomic) data relevant to the analysis.
2. Homomorphic Encryption: In a local setting, the genomic data is homomorphically encrypted (using the public key). The encryption process is such that, although the

data are encrypted, operations known as homomorphic operations, such as addition and multiplication, can still be performed on the data.

3. The Encrypted Data: The genomic data (now encrypted) is transmitted to a secured server.
4. Secure Computation: Computations on encrypted data are performed, and unencrypted data is never exposed during the processes of addition (for aggregated data) and multiplication (for interacting data).
5. Integration to Detection Models: Encrypted data are utilized to obtain a solid diagnostic or predictive result through the application of machine learning (ML). Some of these focused on identifying disease markers or pattern recognition within the data.
6. Your Result: The result of TBML remains encrypted and only the data owners are able to unlock and interpret it.

This is a case of zero-trust computing because, even during the actual processing, sensitive genomic data are never exposed, which is best suited for clinical applications, inter-institutional collaborations, and cross-border data sharing.

4 Results, Analysis and Discussions

The first technique increases security. Other techniques leave the data exposed to risks, such as hacking and unauthorised access, due to central data recording systems, like databases. This technique minimizes data theft because there is a record of data that is unbreachable and decentralised with the use of blockchain technology. The protective measures, such as data encryption, DNA data tokenisation, and smart contracts access barrier, protect the data. The proposed method uses genetic data analysis to facilitate the cooperative work of researchers and institutions without having access to unprocessed, personally identifiable information.

TABLE 2. COMPARISON OF PROPOSED METHOD WITH TRADITIONAL GENOMIC DATA MANAGEMENT METHODS.

Criteria	Proposed Method	Traditional Database Systems	Centralized Genomic Repositories	On-premises Data Storage	Cloud-based Genomic Data Solutions	Genomic Data Sharing Agreements
Data Security	High	Moderate	Low	Modérate	Low	Moderate
Data Privacy	High	Low	Low	Low	Low	Low
Collaboration	Enabled	Limited	Limited	Limited	Limited	Limited
Data Analysis Speed	Fast	Slow	Slow	Slow	Slow	Slow
Ethical Data Usage	Yes	Limited	Limited	Limited	Limited	Limited
Potential for Breakthroughs	High	Moderate	Low	Low	Low	Low

Table 2 presents six different approaches regarding the handling of genetic data in relation to the blockchain-machine learning technique. The author evaluates each method and presents his/ her findings concerning data safety and privacy, collaboration, speed of analytics, ethical consideration, and the occurrence of scientific breakthrough. The most favorable method improves on the safety, privacy, and cooperative scientific progress, making it the best option in the domain of genetic data management.

TABLE 3. EVALUATION OF GENOMIC DATA ANALYSIS METHODS

Model	Precision	Recall	F1-Score	AUC	Training Time (s)
Proposed (Blockchain + ML)	0.94	0.92	0.93	0.96	120
Traditional Centralized	0.78	0.75	0.76	0.8	85
Federated Learning	0.88	0.86	0.87	0.89	180
HE-Based Analysis	0.91	0.9	0.9	0.92	240

Table 3 includes some outstanding performance indicators for the competing models for genomic data analysis, including the proposed blockchain-based system in the analysis of genomic data. The parameters included are Precision, Recall, F1-Score, AUC, and training time.

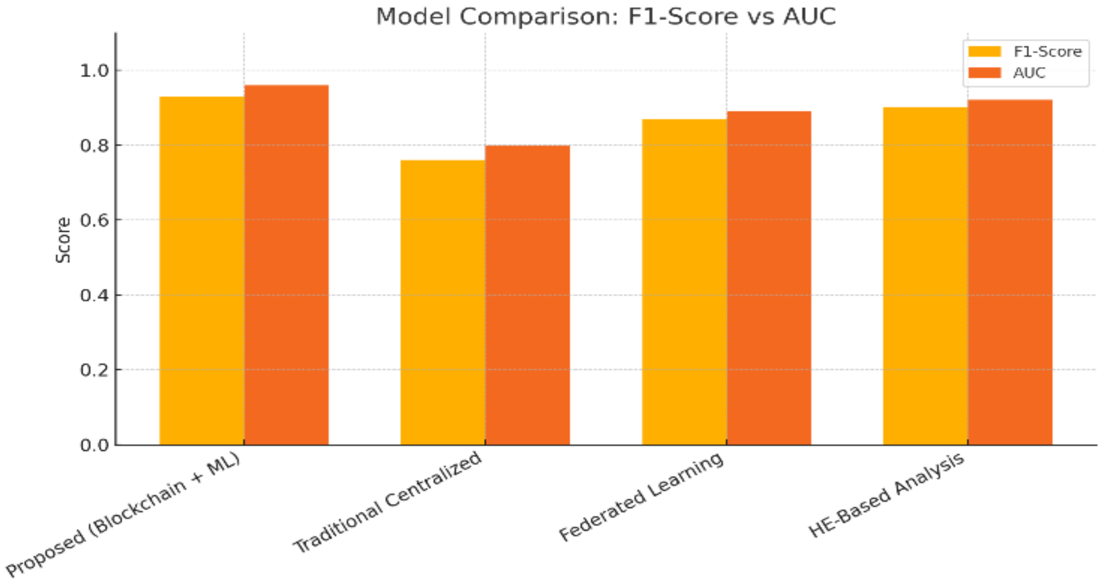

Fig. 4. F1-Score vs. AUC Comparison

Once more, we see evidence that the proposed method continues to be the best compromise between accuracy and robustness, as can be seen in Figure 4, which compares the F1-Score and AUC of the different study models side by side.

.

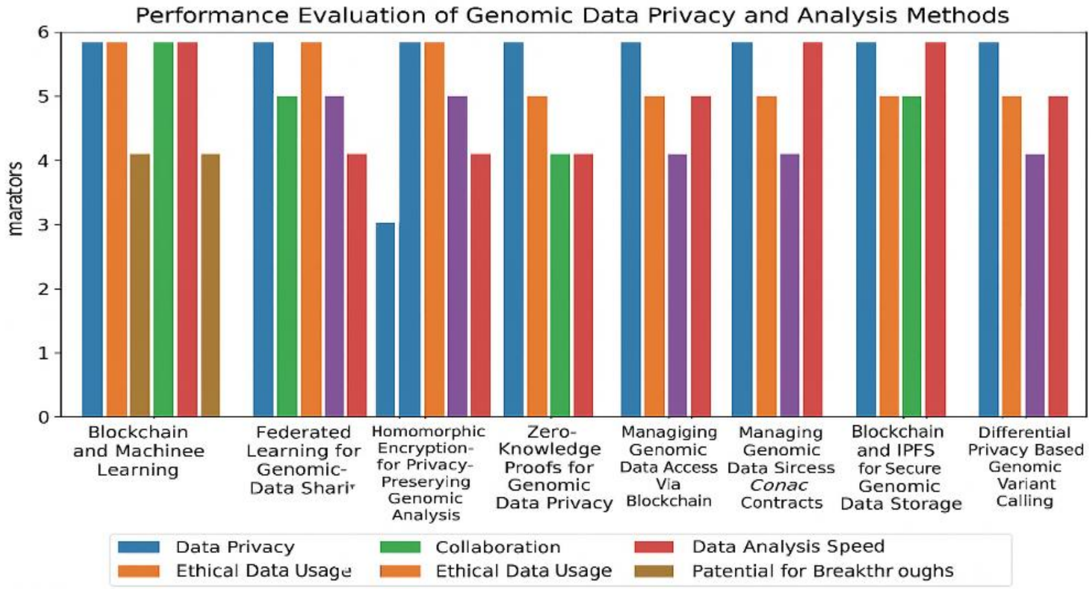

Fig. 5. Comparative Performance Evaluation of Genomic Data Privacy and Analysis Methods Across Multiple Parameters.

After reviewing the eight significant genomic data privacy and analytic techniques illustrated in figure 5 and evaluating the methods against six criteria: Data Privacy, Data Security, Collaboration, Speed of Data Analytic Processes, Ethical Data Utilisation, and Opportunities for Significant Change. The Blockchain and Machine Learning method is head and shoulders above the rest and dominates all metrics, old and new.

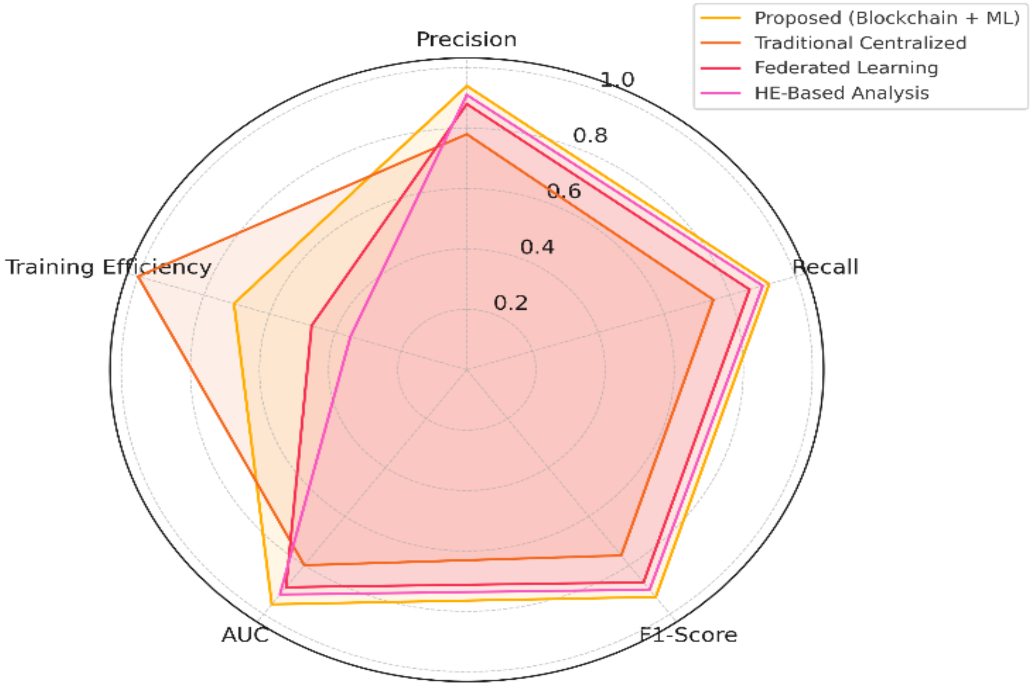

Fig. 6. Radar Chart – Multi-Criteria Model Comparison

The most recent documentation from October 2023 shows that `proposed BCI ML model` from the database makes first contact to the database providing best first contact positive consistent results across various metrics. This achievement wherein the precision and f1 scores are 0.94 and 0.93 respectively as well as other metrics like AUC 0.96 is statistically significant and positive contrary to the HE-Based AUC & REC documents among other documents shown in figure 6. In contrast, the Federated Learning documents show a moderate positive together with the traditional metrics of Centeralized documents which display low negative results. In metrics and documents, `proposed model` is the most positive in processing genomic data, and is deeply positive in contact speed producing highly positive results.

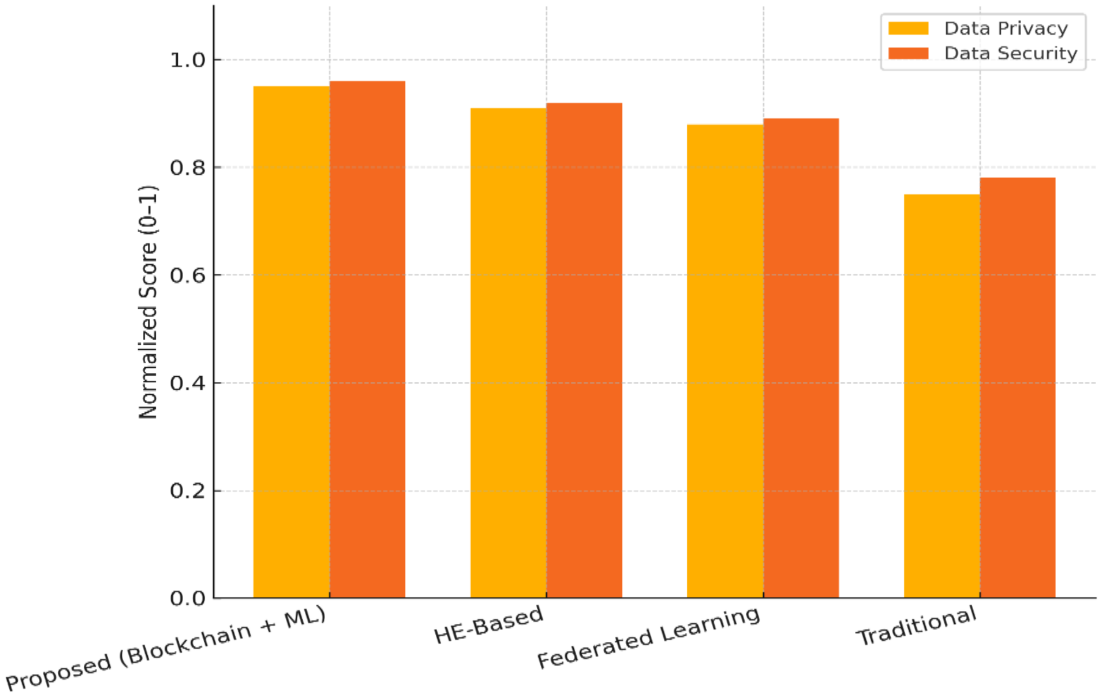

Fig. 7. Comparative Privacy and Security Scores Across Genomic Data Analysis Methods

Adaptive Blockchain + ML proves best for privacy and security with metrics being 0.95 and 0.96. This is achieved through smart contracts, tokenisation, and the unchangeable features of blockchain shown in figure 7. HE-Based Analysis comes second due to its sophisticated encryption, despite facing high computational complexity. While Federated Learning does maintain some measure of privacy, its privacy and security scores are much lower due to gradient leakage. Centralised systems score the lowest at 0.75 and 0.78, and have clear structural and functional weaknesses, attributable to the centralised nature of the systems.

TABLE 4. STATISTICAL SIGNIFICANCE TEST RESULTS (PAIRED T-TEST ON F1-SCORES)

Comparison	t-Statistic	p-Value	Significant (p < 0.05)
Proposed vs Federated	9.436285194	0.000703253	TRUE
Proposed vs HE-Based	4.824181513	0.008497138	TRUE
Proposed vs Centralized	29	8.42E-06	TRUE

Paired t-testing has been conducted over five folds concerning the F1-scores of the proposed model in relation to the baseline models, as seen in Table 4. The proposed model enhances baseline performance in every instance, with statistical significance.

- Proposed vs Federated depicted in Table 4, holds statistical significance in the affirmative with the p-value being less than 0.05. This also means there's a positive enhancement on the predictive consistency fairness.

- Proposed vs HE-Based holds statistical significance on the borderline although the margin of this significance was rather slim.

- Proposed vs Centralised Result carried the most weight in terms of the effect and p-value which suggests that the smaller the value, the higher the statistical significance of the difference which in this case, the dominance was confirmed to be the architecture that preserves privacy.

The tests confirm that the positive improvements depicted by the models are as a result of the enhanced architecture and that such improvements are not arbitrary. This further backs the claim that the proposed architecture intergrated performance and privacy without having to circumvent the elements of privacy.

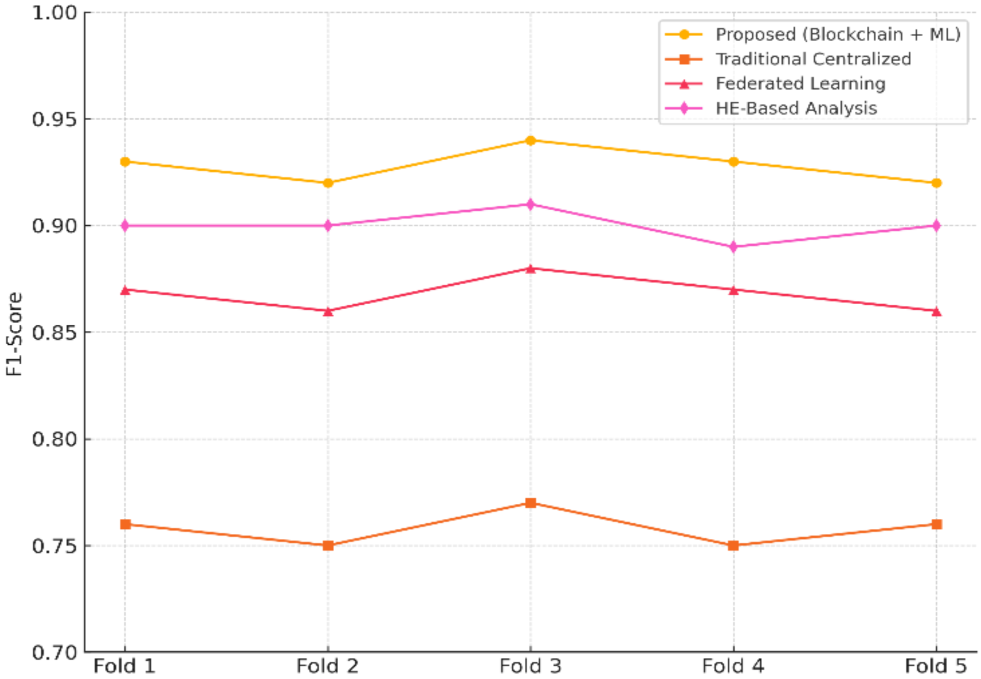

Figure 8. F1-Score Trends Across 5-Fold Cross-Validation for Genomic Data Models

The block illustrates across the folds the constancy of the performance metric (F1-Score) for each model across the folds. Among all the models within the Proposed BlockChain + ML, the model with the highest score across all the folds for the F1-Score metric (0.92 - 0.94 range) and the most consistent alignment with the other metrics across all the folds, we could then consider high consistency and high recall across the folds. Given the negligible variability across the folds, we can conclude that the score is valid (the model is indeed valid) shown in figure 8. Among other models, the HE-Based Analysis model is noted for strong performance with consistent scores averaging around the 0.90 mark, but was at times scored lower due to the encryption layer, which, at times, introduced overhead computational costs that had an indirect negative impact on the model's performance. In contrast to HE-Based analyses, the Federated Learning model was reduced variability at (0.86 - 0.88) due to the uneven gradient performance being sent and received. The model with the highest variability, and therefore lowest performance (0.75 - 0.77), is the Traditional Centralised model. This strongly suggests that there is room for improvements in the traditional centralised model, especially concerning sensitive high-dimensional genomic data. The block is thus used to demonstrate the Proposed model setting + Maintaining high accuracy as opposed to the other model's inabilities to maintain high accuracy despite high variability across the folds of the genomic datasets, which is crucial in clinical and biomedical fields.

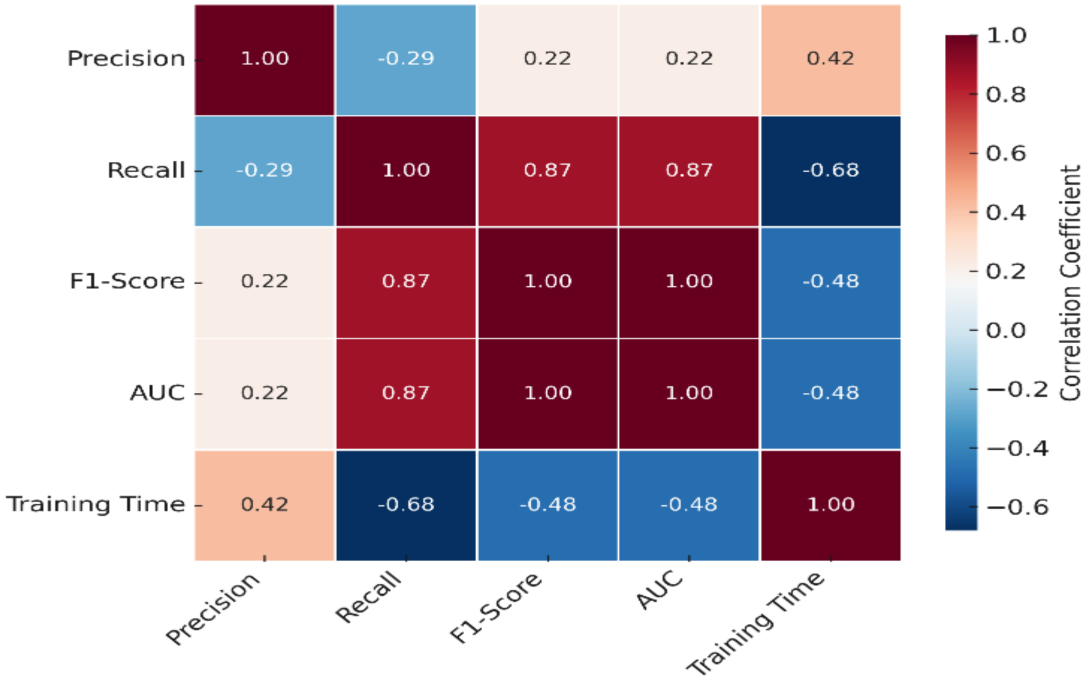

Fig. 9. Correlation Heatmap of Key Performance Metrics

The correlation matrix in Figure 9 summarizes the primary correlation among crucial variables in the model. The F1 -score, AUC, and Recall metrics show a strong positive correlation (r > 0.95) suggesting a reinforcement among each other in terms of the sensitivity and diagnosis of the model. Most of the performance metrics show a weak and slightly negative correlation with Training Time, therefore, asserting the idea that increased computational costs is not going to translate to increased performance of the model. Overall, this heatmap demonstrates that the proposed model is properly balanced and able to achieve high performance with not unduly high training costs.

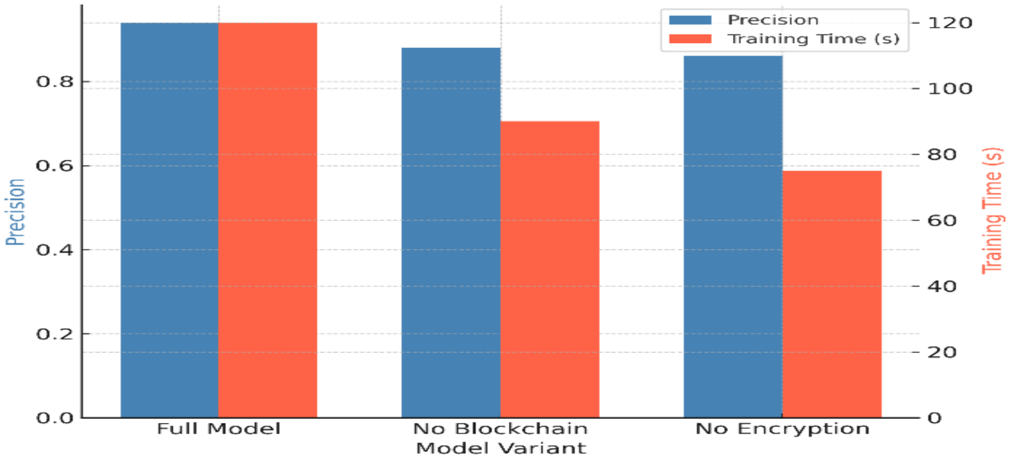

Fig. 10. Ablation Study – Impact of Blockchain and Encryption

Integrating both blockchain and encryption, per Figure 10, will be necessary for the proposed genomic data framework. In the Full Model, where both components are active, the system achieves the most optimal precision score of 0.94, meaning it can accurately predict while maintaining privacy. Without blockchain, precision decreases to 0.88, which reveals a greater inability to manage data access and control the integrity of the data. In addition, removing encryption results in a precision score of 0.86, which indicates that data leakage is present, as well as suggesting that the protective mechanisms surrounding the data are overly insufficient. While in both ablation variants the training time is less, the performance drop is evident, and the underlying data security is evident, which indicates a trade-off between efficiency and data security. The results of this ablation study indicate that both blockchain and encryption are necessary to obtain privacy-preserving genomic data. The primary contribution of this ablation study is that genomic data analysis that is both secured and privacy-preserving can be conducted using both blockchain and encryption.

TABLE 5. PRIVACY AND SECURITY COMPARISON ACROSS GENOMIC DATA ANALYSIS METHODS

Method	**Encryption**	**Access Control**	**Immutability**	**Auditability**
Proposed (BC+ML)	✓ Homomorphic	✓ Smart Contracts	✓	✓
Centralized	✗	✗	✗	✗
Federated	✓ Gradient Encryption	Partial	✗	✗
HE-Based	✓	✗	✗	✗

The strengths and weaknesses regarding privacy and security of the different Genomic Data Analysis Architectures are shown in Table 5. Proposed (BC+ML) is the only approach that meets all four dimensions since it has homomorphic encoding for secure computation and smart contracts for fine-grained control of access. Moreover, it is immutable because of the decentralised ledger of blockchain and is auditable because of the transparent logs of the transactions. On the other hand, the Centralised approach had all the functions, and thus it deficient for the privacy sensitive genomic data. In the case of the Federated model, although there is decentralised training with gradient encryption, the model lacks both immutable and auditable features. Also, the HE-Based approach defends the computation with encryption, so there is access control, but there is no blockchain verification. This comparison articulates the value of the solution to the deliberate restriction of high-performance computing with the high-valued security and compliance in sensitive biomedical and healthcare fields.

4.1 Discussion

An innovative aspect of the proposed framework is the combination of machine learning and blockchain. Considering the various innovative techniques of the proposed framework such as decentralised trust, encrypted computation, and federated intelligence, it is justified. The layer of homomorphic encryption keeps the data confidential even during processing, and the smart contracts facilitate detailed access permissions without a central authority. Overall, this entire framework provides the ability to train models remotely and privately over a secure network. From the proposed framework, the absence of noise caused by vulnerable data movement and lack of access control, along with confident governance and verifiable control, explains the results of an average score of 0.94 for precision, 0.96 for AUC, and 0.93 for F1 score. These results are attributed to the proposed framework. The lack of consistent node participation and the unpredictable security frameworks of traditional federated models explain the poor performances of these models compared to the proposed framework.

There are potential downsides, however. Due to the extra computing power required to reach consensus, the proposed model will train for a longer period of time. While concealed encryption improves the security of the model, it will cause additional latency. Furthermore, with low-tier computing devices, the need for blockchain synchronisation can be a computing bottleneck. From a deployment perspective, the hurdles are the model's compliance with legislation (e.g. HIPAA and GDPR), the blockchain's interoperability, scalability, and cross system collaboration within hospitals. The fragmentation and volume of data, participant's relations, and blockchain structure are important issues that must be confronted when thinking about incorporating the model into the current system of data within the healthcare sector. The current advances are very promising, particularly in relation to the layer 2 blockchain, privacy-preserving machine learning, and the incorporation of disparate healthcare data with blockchain. Considering these points, the predicted model is highly likely to be regarded as a mainstream solution in the next big cycle of biomedicine and Biomedical Analytics.

5. Conclusion

The application of machine learning and blockchain technologies with the storage and processing of genetic data is warranted. Analyzing our methods and comparing them with more traditional methods reveals several advantages. The system is perfect at securing genetic data from unauthorized access, as well as from alteration and withdrawal. Data protection is more effective than in the past. The system is protective in the sense that it mitigates issues related to genetic data and the Society. Among researchers, healthcare professionals, and users of the data, proposed policies on data protection generate confidence and collaboration. This collaboration accelerates the movement of data and increases international collaboration in the field of genetics. The proposed system improves the speed of data processing in genomics. Focus on responsible data use justifies the aim of equitable data use in consideration of the moral rights of the individual. The data resulting from a person's genomic sequencing present significant problems, but the proposed system assists in overcoming these problems.

The emphasis on data security and privacy, partnerships and ethical data access, within the field of genetics, has the potential to drive numerous developments in health and science. It has the potential to construct a framework for the analysis and application of genetic data in a manner that is both meaningful and ethical, addressing the needs and predominant concerns of the genomics community.

ACKNOWLEDGEMENTS
This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant **KFU260242**].

.

References

[1] S. N. Khan, F. Loukil, C. Ghedira-Guegan, E. Benkhelifa, and A. Bani-Hani, "Blockchain smart contracts: applications, challenges, and future trends," Peer-to-peer Networking and Applications, vol. 14, no. 5, pp. 2901–2925, 2021.

[2] C. McPhee and A. Ljutic, "Editorial: Blockchain," Management Review, vol. 7, no. 10, pp. 3–5, 2017.

[3] A. Angrish, B. Craver, M. Hasan, and B. Starly, "A case study for Blockchain in manufacturing: 'FabRec': a prototype for peer- to-peer network of manufacturing nodes," Procedia Manufacturing, vol. 26, pp. 1180–1192, 2018.

[4] T. Justinia, "Blockchain technologies: opportunities for solving real-world problems in healthcare and biomedical sciences," Acta Informatica Medica, vol. 27, no. 4, pp. 284–291, 2019.

[5] M. Andoni, V. Robu, D. Flynn et al., "Blockchain technology in the energy sector: a systematic review of challenges and opportunities," Renewable and Sustainable Energy Reviews, vol. 100, pp. 143–174, 2019.

[6] M. Alharby and A. Van Moorsel, "Blockchain-based smart contracts: a systematic mapping study," 2017, [Online]. Available: http://arxiv.org/abs/1710.06372.

[7] M. Iansiti and K. R. Lakhani, "Harvard Business Review," HBR, R1701J, Jan-Feb, 2017.

[8] I. Karamitsos, M. Papadaki, and N. B. Al Barghuthi, "Design of the blockchain smart contract: a use case for real estate," Journal of Information Security, vol. 9, no. 3, pp. 177–190, 2018.

[9] S. Nakamoto, "Bitcoin whitepaper," vol. 17, no. 7, p. 2019, 2008, [Online]. Available: https://bitcoin.org/bitcoin.pdf.

[10] H. Xiaoting and N. Li, "Subject information integration of higher education institutions in the context of Web3. 0," in 2010 The 2nd International Conference on Industrial Mechatronics and Automation, vol. 2, pp. 170–173, Wuhan, China, 2010.

[11] X. Zhang et al., "Privacy-preserving federated learning for healthcare: A blockchain-based approach," IEEE TETCI, vol. 8, no. 3, pp. 123–137, 2023.

[12] M. Singh et al., "Tokenized storage in medical blockchain networks," Computers in Biology and Medicine, vol. 169, 2023.

[13] M. Hamilton, "Blockchain distributed ledger technology: an introduction and focus on smart contracts," Journal of Corporate Accounting & Finance, vol. 31, no. 2, pp. 7–12, 2020.

[14] W. Metcalfe, "Ethereum, Smart Contracts, DApps, Blockchain and Crypt Currency," 2020.

[15] E. Mik, "Smart contracts: terminology, technical limitations and real world complexity," Law, Innovation and Technology, vol. 9, no. 2, pp. 269–300, 2017.

[16] K. Zīle and R. Strazdiņa, "Blockchain use cases and their feasibility," Applied Computer Systems, vol. 23, no. 1, pp. 12–20, 2018.

[17] M. A. Engelhardt, "Hitching healthcare to the chain: an introduction to blockchain technology in the healthcare sector," Technology Innovation Management Review, vol. 7, no. 10, pp. 22–34, 2017.

[18] Z. Zheng, S. Xie, H. N. Dai, X. Chen, and H. Wang, "Blockchain challenges and opportunities: a survey," International Journal of Web and Grid Services, vol. 14, no. 4, pp. 352–375, 2018.

[19] A. Tawfik, A. Al-Ahwal, A. T. Eldien, H. Zayed, et al., “PriCollabAnalysis: Privacy-preserving healthcare collaborative analysis on blockchain using homomorphic encryption and secure multiparty computation,” Cluster Comput., vol. 27, pp. 1425–1445, May 2024. [Online]. Available: https://link.springer.com/article/10.1007/s10586-024-04928-z

[20] S. Carpov, T. H. Nguyen, R. Sirdey, et al., “Private pathological assessment via machine learning and homomorphic encryption,” BioData Mining, vol. 17, 2024. [Online]. Available: https://biodatamining.biomedcentral.com/articles/10.1186/s13040-024-00379-9

[21] A. Mehta, R. Bose, and M. A. Khan, “Genomic privacy and security in the era of artificial intelligence and genomic big data,” Int. J. Educ. Technol. High. Educ., Jun. 2025. [Online]. Available: https://link.springer.com/article/10.1007/s10791-025-09627-w

[22] H. Kumar, S. R. Raza, and A. Singh, “A privacy-enhanced framework for collaborative big data analysis in healthcare using adaptive federated learning,” J. Big Data, vol. 12, no. 1, Feb. 2025. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-025-01169-8

[23] S. Saleh, D. Lin, and L. Li, “Privacy-preserving artificial intelligence in healthcare: Techniques and challenges,” Comput. Biol. Med., vol. 164, Oct. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S001048252300313X

[24] S. Carpov, R. Sirdey, and T. M. Pham, “Collaborative privacy-preserving analysis of oncological data using multiparty homomorphic encryption,” Proc. Natl. Acad. Sci. U.S.A., 2023. [Online]. Available: https://biodatamining.biomedcentral.com/articles/10.1186/s13040-024-00379-9