GMR
The Original

PGCA FL: A PATHWAY CONSENSUS-GUIDED FEDERATED LEARNING FRAMEWORK FOR PRIVACY-PRESERVING DISTRIBUTED GENOMIC CANCER CLASSIFICATION

M. Shobana[1*], V. Sathya Preiya[2], S. Suresh[3], A. Kanchana[4], V. Gowri[5]

[1*]Assistant Professor, Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, Orcid id: https://orcid.org/0009-0000-7090-1611, Email: shobanamurugan@panimalar.ac.in
[2]Associate Professor, Department of Computer Science and Engineering, S.A. Engineering College, Chennai, https://orcid.org/0000-0002-6203-1289, Email: sathyapreiya@yahoo.com
[3]Assistant Professor, Department of Artificial Intelligence and Data Science, Panimalar Engineering College, Chennai, Orcid id: https://orchid.org/0000-0002-6429-9879, Email: beintellectuals@gmail.com
[4]Assistant Professor, Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, Orcid id: https://orchid.org/0000-0002-1025-9046, Email: akanchana2683@gmail.com
[5]Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Orcid id: https://orcid.org/0000-0003-2303-1101, Email: gowriv2@srmist.edu.in

ABSTRACT

The growing amount of genomic data has accelerated precision medicine by making it possible to analyze genomic data through artificial intelligence (AI) and classify cancers by their genomes. Yet, there are privacy regulations, institutional ownership of data and restricted access to distributed health resources issues that detract from centralized genomic learning. While federated learning (FL) offers a potential solution, the current federated learning paradigms are primarily statistical, and lack the incorporation of biological relationships between genes and pathways. This study introduces a privacy-preserving approach to multi-institution genomic cancer classification, called PCGA-FL (Pathway Consensus-Guided Federated Learning), that incorporates biological pathway knowledge into federated optimization. This proposed framework features a Pathway Consensus Index (PCI) to measure pathway-level agreement across clients, a Pathway-Aware Local Representation Learning (PLRL) approach to extract biologically meaningful genomic features and a Consensus-Guided Federated Aggregation (CGFA) algorithm for adaptive global model optimization. The framework is tested with the publicly available Cancer Genome Atlas (TCGA) RNA-sequencing (RNAseq) dataset for multiple cancer types, where a multi-institutional federated learning environment is emulated by splitting genomic data across a number of institutional clients. Experimental results show that PCGA-FL gets 98% accuracy, 94% precision, 93% recall, 93.5% F1 score, and 0.96 AUC, and enhances the convergence efficiency and communication overhead reduction of conventional federated learning methods. The proposed framework illustrates how to combine the power of consensus between biological pathways and privacy-preserving federated learning for privacy-preserving and interpretable cancer classification in a simulated multi-institutional federated learning setting.

KEYWORDS: Federated Learning, Genomic Cancer Classification, Pathway Consensus, Privacy Preservation, Precision Medicine and Distributed Artificial Intelligence.

1. INTRODUCTION

With this speed of progress in sequencing technologies and the availability of vast genomic resources, precision medicine has been opened to approaches via artificial intelligence (AI) that enable researchers to detect complex patterns between genomes and disease outcomes. There has been remarkable success in using machine learning and deep learning models to identify high-dimensional molecular features and aid in the discovery of biomarkers for disease prediction in the genome. The Cancer Genome Atlas (TCGA) Pan-Cancer Analysis Project developed a comprehensive genomic resource to facilitate, for the first time, large-scale studies of molecular changes in a variety of cancer types. Additional studies have further shown that adding genomic data to other biological data measurements can enhance the prediction ability for the disease. The Nightingale Health Biobank Collaborative Group performed analysis on the genome and metabolome profiles of 700,217 individuals, which proved the ability of molecular integration for common disease prediction [2]. Garg et al. also demonstrated the utility of multi-omics and biomarker-based analyses for improving genetic discovery using UK Biobank data [3].

The advancement of genomic analysis has also been enhanced by deep learning with the ability to learn non-linear relationships between molecular features. Pathomic Fusion used multimodal deep learning for cancer diagnosis and prediction of prognosis using histopathological and genomic data [1]. DeepRisk showed the potential of deep neural architectures in genome wide disease risk assessment [5]. Comparative evaluation of machine learning techniques for prediction of genetic diseases has also pointed out the significance of computational intelligence in classification problems in genome [4]. These strategies, however, are typically based on centralized processing of the data, thus hindering their

implementation in real healthcare settings where genomic data is spread across hospitals and research centers as a result of privacy, ethical and ownership considerations.

Federated learning (FL) is a potential answer as it allows federated institutions to train models together without sharing patients' raw data. Prototype-based aggregation, adaptive optimization, robust training and privacy-aware learning strategies have been explored in recent few years in the field of FL under the context of distributed medical intelligence [6]–[14]. Current FL methods however mostly aggregate based on the statistical properties of updates from local models and neglect the biologic relations between genes and functional pathways. This is important for genomic applications since disease mechanisms are frequently tied to coordinated interactions between many genes, rather than independent, identifiable genomic features.

Biological pathway-based learning offers a chance to consider domain knowledge in the context of genomic prediction and involve groups of functionally related genes. The use of pathway aware modeling has been shown in previous studies to enhance the interpretation of the genome and the predictive performance of machine learning approaches thanks to the fact that biological relationships are taken into account [22]–[24]. However, the pathway-based approaches are primarily developed for centralized systems, and the federated learning approaches typically have no biological awareness when optimizing the system. Thus, there is a need to develop a unified framework that integrates pathway knowledge with privacy-preserving federated learning, which remains a research challenge.

To fill this gap, this research introduces privacy-preserving framework, namely PCGA-FL (Pathway Consensus-Guided Federated Learning), in genomic disease prediction in a distributed institutional learning environment. Unlike traditional FL methods solely based on parametric aggregation, PCGA-FL-adds the concept of pathway consensus in federated optimization by incorporating the notion of biological agreement among the participating clients. The proposed framework combines the pathway-aware representation learning, a Pathway Consensus Index (PCI), and consensus-guided aggregation, providing biologically informed genomic intelligence without compromising the data privacy. PCGA-FL adopts pathway-level biological consensus as an optimization criterion for genomic model aggregation, which is different from the current federated learning frameworks that only optimize client similarity with statistical and/or model-based information.

This study has the following main contributions:

- The proposed work in this project aims to develop an idea for privacy-preserving multi-institution genomic cancer classification, by incorporating biological pathway knowledge via Pathway Consensus-Guided Federated Learning (PCGA-FL) in federated optimization.
- A Pathway-Aware Local Representation Learning (PLRL) mechanism is proposed to learn biologically meaningful genomic patterns and a Pathway Consensus Index (PCI) is proposed to measure the consensus among distributed clients regarding the learned pathways.
- A Consensus-Guided Federated Aggregation (CGFA) strategy is designed to optimise the model update in a global federated learning system based on the pathway consensus information, while ensuring learning stability when the genomic distributions are heterogeneous.
- It is validated on a simulated federated learning setting with the TCGA RNA-sequencing dataset for multi-class cancer classification, where the proposed framework is shown to yield superior prediction accuracy, convergence, and communication efficiency over previous research.

2. RELATED WORK

2.1 Federated Learning and Privacy-Preserving Healthcare Intelligence

Federated learning (FL) is a promising solution to collaborative healthcare intelligence without sharing raw patient information, and it has prompted several institutions to work together to create machine learning models. Current works in the field of healthcare FL have mainly addressed issues of aggregation strategy, dealing with data distribution heterogeneity, and privacy in distributed environments.

Under the scenario of medical imaging data from multiple sources with various imaging environments, Gao et al. designed a prototype-based clustered federated learning framework FedPC which uses prototype information to enhance the robustness of federated learning under heterogeneous medical imaging environments [6]. To enhance the generalization ability over distributed medical segmentation clients, Jin et al. proposed an adaptive multi-stage federated learning (MSAFed) approach [7]. FL with the help of blockchain technology, explainable FL, robust decentralized FL, and asynchronous optimization scheme for healthcare applications have been explored in other works [8]–[13]. Although these strategies show the potential of distributed healthcare intelligence, these are predominantly imaging and clinical applications and do not take into account the biological nature of genomic data.

The preservation of privacy has always been one of the major issues in healthcare FL due to the sensitivity of medical and genomic data. There are several studies that have explored secure aggregation, cryptographic protection and robust optimisation mechanisms, to enhance the confidentiality and reliability. Peng et al. proposed a privacy-preserving aggregation scheme for MD data by secure computation techniques [15]. To enhance the privacy-preserving federated learning system against malicious updates, Dong et al. proposed a Byzantine-robust privacy-preserving federated learning framework [16]. Tanuwidjaja et al., surveyed the privacy-preserving deep learning methods based on encryption, secure computation and privacy-enhancing mechanisms [17]. Access control mechanisms for the protection of healthcare data based on blockchain and encrypted learning and secure aggregation methods have also been explored in the literature [18]-[20]. Wang et al. investigated classification in multiple cloud-based healthcare platforms with privacy preservation to realize collaborative analysis in a secure manner [21].

These methods enhance the privacy and security of the distributed learning, but they primarily target the protection of information exchange and the protection of communication. Typical FL aggregation methods involve counting the number

of updates within each local neighborhood and are not based on any biological information about the genes, molecular functions or pathways that are associated with a disease. This restriction is the fact that this is applicable to a limited use cases like genomic disease prediction, where biological knowledge is important.

2.2 Pathway Based Genomic Learning and Research Gap

Biologically meaningful is the approach of the pathway-based genomic learning that represents groups of functionally related genes related to the disease mechanisms. Pathway-aware approaches build on the foundation of pathway-based prior biological knowledge to enhance the interpretability of such models and to gain insight into disease-associated molecular processes.

Wang et al. illustrated the use of pathway-level representations for genomic classification and analysing the interactions between genes, which were also achieved using machine learning models, and the incorporation of biological functional relationships [22]. Boluki et al. added the pathway information into predictive models and demonstrated that adding biological information could increase the prediction capability and enhance the interpretability of the models [23]. Bao et al. developed an enhanced pathway impact analysis method based on gene weighting approaches, called gwSPIA, to discover signaling pathways associated with a disease [24]. Recently, precision medicine research highlighted the need to combine information from multiple omics with AI analysis for deeper insights into the biomedical field [25].

Concurrently, AI researchers have emphasized the need for transparency, reliability, and responsible use of AI in health care systems. Federated explainable AI systems have shown the potential benefits of an interpretable distributed model [26] and wider research has considered how to be responsible, equitable and trustworthy in the evaluation of medical and genetic prediction systems [27]–[29]. These approaches, however, tend to be more focused on explainability and reliability, while not considering biological pathway knowledge in federated genomic optimization.

In this context, previous research focuses on three areas: genomic prediction, federated learning, privacy preservation, and biological interpretation. The current federated genomic prediction methods do not consider pathway-level consensus as an optimization criterion for the model aggregation. There is limited availability of a unified framework that integrates pathway-level biological knowledge with federated optimization techniques that are privacy preserving. The proposed PCGA-FL framework tackles this research gap by introducing the new concept of a pathway consensus guided optimization that allows for distributed genomic learning, taking into account the agreement between different distributed institutional clients.

The representative approaches related to genomic AI, federated learning in healthcare, and pathway-based analysis are summarized in Table 1. Existing techniques are able to solve individual issues well but integrating pathway knowledge with federated genomic learning with privacy is still underdeveloped.

Table 1. Comparison of Representative Existing Approaches

Study	Domain	Technique	Main Limitation
Pathomic Fusion [1]	Cancer prediction	Multimodal genomic AI	Centralized data dependency
DeepRisk [5]	Disease prediction	Deep genomic learning	No privacy-preserving collaboration
FedPC [6]	Healthcare FL	Prototype-based FL	No pathway information
Tian et al. [11]	Healthcare FL	Robust privacy FL	No biological guidance
Wang et al. [22]	Genomic analysis	Pathway-based learning	No federated optimization
Boluki et al. [23]	Biomedical prediction	Pathway-guided modeling	Centralized framework
PCGA-FL (Proposed)	Genomic disease prediction	Pathway consensus-guided FL	Integrates biological knowledge with privacy-preserving collaboration

PCGA-FL (Proposed) Genomic disease prediction Pathway consensus-guided FL Integrates biological knowledge with privacy-preserving collaboration While these approaches are effective to use FL in healthcare, their corresponding optimization strategies are typically designed for imaging and clinical modalities, and are not directly applicable for optimizing FL for genomic data, which are inherently biological dependent.

3. PROPOSED PCGA-FL FRAMEWORK

3.1 Framework Overview

In this study, we introduce a privacy-preserving federated learning framework, called PCGA-FL (Pathway Consensus-Guided Federated Learning), that predicts genomic diseases in decentralized healthcare systems. PCGA-FL aims to overcome this weakness by optimizing a global model primarily based on statistical characteristics of local data distributions while taking into account biological relationships (interactions) between genomic features.

For disease prediction at the genomic level, the molecular features that can correlate with disease outcomes can be expressed as interactions between different genes that form functional biological pathways. But traditional federated learning methods consider genomic features as independent variables, and fails to make use of the prior knowledge of the pathway in the collaborative optimization process. Thus, the learned global model can be successfully used for prediction without being biologically consistent at distributed clients.

For this reason, PCGA-FL adds a pathway-guided optimization process to the federated learning process. The framework allows for the learning of pathway-aware genomic representations from locally store data from each client, and the

generation of pathway importance profiles that reveal the biological patterns associated with diseases. These profiles are used by the federated server to calculate the level of agreement over the pathways that the clients will be using, and can be used to fine-tune the weights of the models used for aggregation.

The proposed approach introduces two novel features: Pathway-Aware Local Representation Learning (PLRL) to incorporate genomic features with biological pathway information into the client-side training and Consensus-Guided Federated Aggregation (CGFA) to incorporate pathway consensus into the global model optimization. With this design, PCGA-FL supports collaboration on genomic data without compromising privacy, enhances biological interpretability, and allows for consistency across different clients.

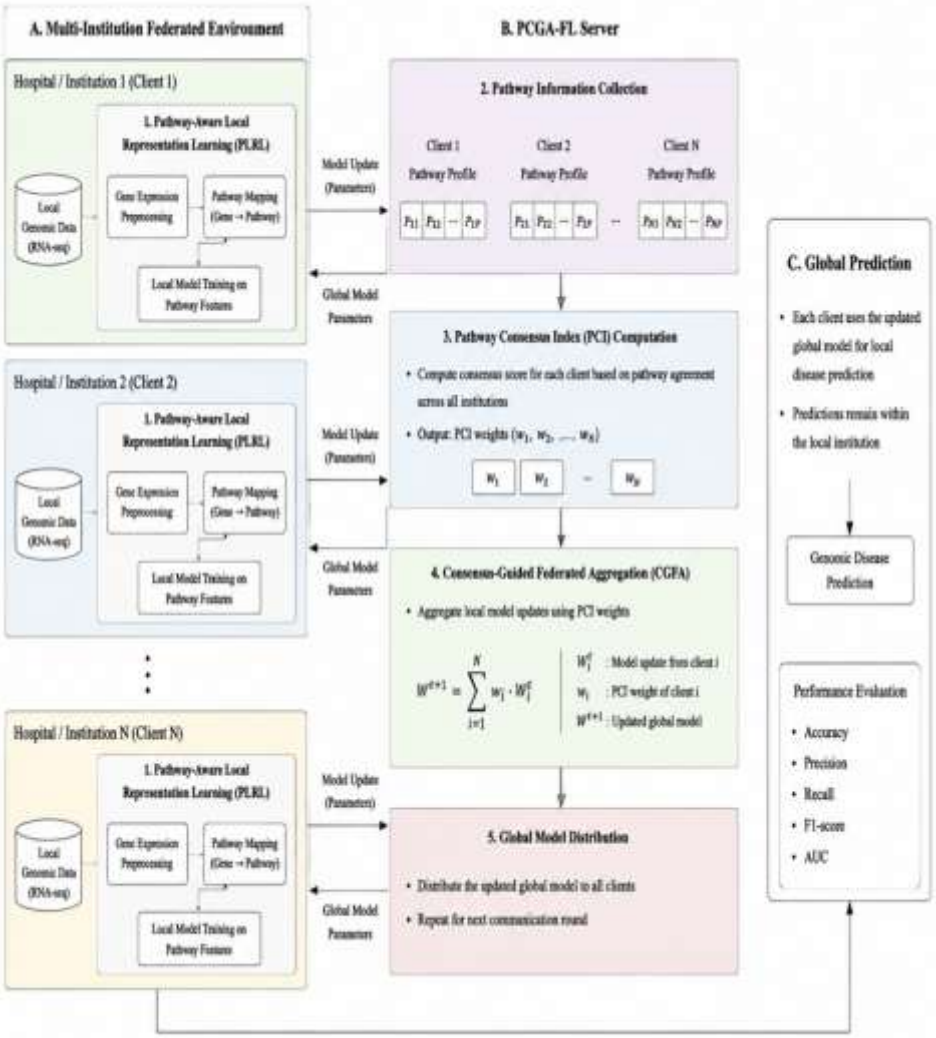

Figure 1. Overall architecture of the proposed PCGA-FL framework for privacy-preserving multi-institution genomic cancer classification.

The project spreads genomic samples across a number of institutional clients. The representations of each pathway and model parameters with pathway importance information are generated independently for each client. Pathway consensus is computed on the server, and an adaptive aggregation of pathways is implemented without the need to access the raw genomic data.

3.2 Dataset Description and Federated Environment Construction

The proposed framework is tested with the available genomic information from The Cancer Genome Atlas (TCGA) Pan-Cancer Analysis Project [30] data which is publicly available. TCGA offers large-scale molecular data profiles such as RNA-sequencing expression data and clinical annotations for several cancer cohorts, paving the way for computational analysis of disease-associated molecular patterns.

A distributed healthcare environment is created by dividing the TCGA genomic data into several institutional clients where the data is typically fragmented and cannot be used in its entirety in the real world because of privacy and ethical issues. Every client is an independent healthcare organisation with its own local genomic data, and engaged in federated training, with no sharing of raw patient-level data.

The whole data set is given by:

$$D = \{D_1, D_2, \ldots, D_K\} \qquad (1)$$

The number of participating institutional clients is represented by where (K). All clients are responsible for local pathway-aware learning, and the central server optimizes the model based on the information of only the model parameters transmitted from the clients and the pathway information.
This experimental design allows for the evaluation of PCGA-FL in realistic FL setting while maintaining genomic healthcare application privacy requirements.

3.3 Pathway-Aware Local Representation Learning
In the first step of PCGA-FL, we will be working to extract biologically meaningful genomic representations at each client. The proposed method extends the conventional federated model which optimizes directly on the level of gene-level feature vectors by adding pathway information to reflect functional relationships between genes.
The genomic representation is improved for each client (k) by incorporating learned genomic features and pathway embeddings:

$$H_k = \phi(W_k X_k) + \lambda\psi(P_k) \qquad (2)$$

The pathway-aware genomic representation is denoted as Hk, the local model parameters are denoted as Wk, the genomic feature learning is denoted by (ϕ) and the pathway embedding extraction is denoted by ψ(Pk). This is because it factors in pathway information in learning the local representation this is controlled by the parameter λ.
The representation generated is a pathway-aware representation which allows each client model to represent predictive genomic characteristics and biologically associated molecular patterns. PLRL offers a different approach to FL than the existing methods in healthcare that optimize statistical representations, by adding biological context before federated aggregation.
The Biological Agreement Measurement (BAM) is characterized by a Pathway Consensus Index. There is a Pathway Consensus Index for the Biological Agreement Measurement (BAM). PCGA-FL's biggest achievement is a Pathway Consensus Index (PCI) that measures the level of agreement among distributed clients in the biological realm. Following local training, clients each create a pathway importance profile:

$$V_k = [v_1, v_2, \dots, v_q] \qquad (3)$$

In this case each element is the contribution of pathway q to the local prediction process.
Pathway profile alignment is used to measure the similarity between two clients:

$$S_{ij} = \frac{V_i \cdot V_j}{\|V_i\|\|V_j\|} \qquad (3)$$

where Sij is the similarity of the pathways between client i and j.
The pathway consensus value for client k is obtained from the following equation:

$$PCI_k = \frac{1}{K-1}\sum_{\substack{j=1 \\ j\neq k}}^{K} S_{kj} \qquad (4)$$

The PCI value gives an indication of the match between a client's learned biological patterns and the patterns of the other clients participating. PCGA-FL is different from traditional federated optimization, which only considers data contribution as a weight of the clients.

3.5 Consensus-Guided Federated Aggregation
Most of the federated learning algorithms currently in use (e.g., FedAvg) are based on the size of the client data. Such aggregation is appropriate for general use, but it is not taking into account if the locally learnt genomic patterns are biologically consistent.
To mitigate this constraint, PCGA-FL proposes a consensus-driven aggregatory approach based on consensus on the linkage and representation of local data. The proposed change for Worldwide is:

$$W^{t+1} = \sum_{k=1}^{K} \alpha_k W_k^t \qquad (5)$$

The local model parameters at round t for communication Wkt, and the adaptive aggregation coefficient αk.
The aggregation coefficient is a number that is defined as:

$$\alpha_k = \frac{N_k PCI_k}{\sum_{j=1}^{K} N_j PCI_j} \qquad (6)$$

where Nk is a sample contribution for the local sample and PCIk is a pathway consensus. The formulation allows the global model to consider the clients who are able to offer enough genomic information and the biologically consistent pathway patterns.

3.6 PCGA-FL Training Algorithm
Algorithm 1. Pathway Consensus-Guided Federated Learning
Input:
The distributed genomic datasets (D1, D2, ..., DK)
pathway information P,
initial model parameters W0
Number of rounds for a communication problem T
Output:

optimized global model W*
Set the initial value of the global model parameters (W0).
For each iteration round t of the communications:
The server broadcasts Wt to all the clients.
Each client performs:
1. Pathway-aware representation learning
2. Local model optimization
3. Pathway importance estimation
Clients transmit:
- Locally determined model parameters Wk
- Pathway profile Vk
The Server can measure the pathway consensus PCI.
Calculate adaptive aggregation weights, αk.
Update global model:

$$W^{t+1} = \sum_{k=1}^{K} \alpha_k W_k^t$$

Return final global model W*.
An explanation of the importance of preserving privacy and considerations for computation.
By keeping raw patient-level genomic data on each institutional client, PCGA-FL ensures that each client has their genomic privacy preserved. The central server does not read or write clinical information or gene expression profiles, but only communicates with the central server the model parameters and pathway level information that is needed for the optimization.
Pathway similarity calculation and consensus estimation are the main costs of PCGA-FL for the extra computations. Pathway profiles still have relatively small dimensionality compared to raw genomic features, and have the benefit of being aggregated from a biological perspective.
The proposed PCGA-FL framework is an extension of the conventional federated learning into distributed genomic optimization, that utilizes the Pathway information provided by Reactome pathway database. The framework facilitates genome-wide data sharing amongst distributed clients with biological consistency and privacy-preserving collaborative representation learning via pathway-aware representation learning and consensus aggregation.

4. EXPERIMENTAL SETUP

4.1 Dataset and Federated Evaluation Protocol

A federated evaluation protocol and a dataset will be developed. The proposed PCGA-FL framework is assessed by publicly available The Cancer Genome Atlas (TCGA) Pan-Cancer Analysis Project (PCA) genomic resource [30]. The data set includes expression data from RNA-sequencing assays and clinical information from several cancer cohorts which can be used to computationally analyze the genomic patterns associated with disease prediction.
The genomic input features in this study are the TCGA RNA-sequencing profiles and the biological pathway information is fed for pathway-aware representation learning. Because of privacy concerns, real-world genomic datasets are kept in separate healthcare institutions and cannot be shared directly, a simulated multi-institution federated environment is created by dividing the genomic samples among the multiple clients. TCGA RNA-sequencing expression profiles are used for multi-class cancer type classification. The gene expression profile for each sample is used as a representation of the sample and the corresponding cancer cohort label is used as a prediction target. For this study, the selected TCGA cancer cohorts serve as the basis to create a heterogeneous genomic classification environment, which represents distributed health care institutions.

Table 2. Characteristics of the TCGA RNA-Sequencing Dataset Used for Multi-Class Cancer Classification

Parameter	Value
Dataset	TCGA RNA-seq
Task	Multi-class cancer classification
Number of samples	10,000
Number of cancer classes	5
Number of genes/features	500
Number of clients	5

Local model training is done by each client using their own genomic samples and only parameters of the model and information about the pathways are communicated to the central server. This setting will allow to test PCGA-FL under privacy-preserving distributed genomic learning setup.

4.2 Implementation Configuration and Comparative Methods

The proposed framework is realized in a federated deep learning system, consisting of multiple client models that optimizes a global prediction model in a round-robin fashion. The most important experimental setups are briefly summarized in Table 2.

Table 3. Experimental Configuration

Parameter	Setting
Dataset	TCGA RNA-seq profiles
Simulated clients	5 institutional clients
Communication rounds	100
Local epochs	10
Optimizer	Adam
Learning rate	0.001
Batch size	32
Aggregation	PCGA-FL consensus aggregation

PCGA-FL is benchmarked against representative algorithms, FedAvg, FedProx, which is a baseline for federated learning without compromising privacy, and a centralized pathway-based genomic model. The first two approaches are examples of conventional federated optimization, heterogeneous federated learning, and privacy-aware learning, respectively, while the final approach is a biological knowledge-based modeling approach. In the ablation experiments, the pathway-aware representation learning (PLRL), pathway consensus estimation (PCI), and consensus-guided aggregation (CGFA) are ablated to study their contribution.

4.3 Evaluation Criteria

The proposed framework is assessed based on the prediction performance, convergence behavior and communication efficiency. The accuracy, precision, recall, F1-score and AUC are used to measure classification performance. Convergence analysis is carried out through the observation of model performances over the communication rounds and communication efficiency is considered by observing the necessary information exchange requirements compared to the federated approaches.

The compared models in the experiments are trained under the same experimental conditions and multiple training runs are conducted to guarantee the reliability of the estimation of the models' performance. Observed improvements are statistically significant analyzed.

5. RESULTS AND DISCUSSION

The proposed PCGA-FL framework was tested and compared to the performance of the traditional federated learning and pathway-based genomic learning models on the standard RNA-sequencing data from the TCGA project. The results are presented in comparative terms in terms of accuracy, precision, recall, F1 score and AUC in Table 3. The proposed PCGA-FL framework is better than all the baseline approaches in terms of better predictions. The model obtains an accuracy of 98.0%, precision of 94.0%, recall of 93.0%, F1-score of 93.5%, and AUC of 0.96. These enhancements illustrate the positive impact that incorporating knowledge from biological pathways into federated optimization can have.

While conventional federated learning methods like FedAvg and FedProx offer good performance, they base their aggregation strategy primarily on statistical updates to the models without taking the biological relationships between the genomic features into account. Thus, local models can learn different representations because of differences in the genomic distributions of the clients.

PCGA-FL, on the other hand, endows the global model with pathway-aware representation learning and pathway consensus-guided aggregation, allowing the global model to focus on the updates to the client that show higher biological agreement. The results of the AUC gain results show that the proposed framework demonstrates superior ability to discriminate between normal and abnormal in capturing the consistent patterns of the disease associated with the genomic information distributed across multiple clients.

Table 4. Prediction Performance Comparison of PCGA-FL with Baseline Methods

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
FedAvg	91.8	88.5	87.6	88.0	0.91
FedProx	93.1	90.2	89.7	89.9	0.92
Privacy-Preserving FL	94.2	91.4	90.5	90.9	0.93
Centralized Pathway Model	95.4	92.1	91.7	91.9	0.94
PCGA-FL (Proposed)	98.0	94.0	93.0	93.5	0.96

5.2 Convergence Behaviour and Communication Efficiency Analysis

All results reported are the average results of several independent training runs. For PCGA-FL, its behaviour is investigated by observing the performance of the various models over the communication rounds. The convergence trend of the proposed framework is shown in Figure 2 as compared to the conventional federated learning approaches.

By integrating the pathway consensus information while aggregating, PCGA-FL can have faster convergence. The conventional Federated Learning method involves statistical contribution to the client updates, and can cause optimization instability if there are differences in the genomic frequency distribution between the clients. The problem is especially significant for genomic applications, where there can be diverse molecular properties in various institutions.

This variation can be mitigated with the proposed Consensus-Guided Federated Aggregation (CGFA) mechanism that assigns higher aggregation importance to clients that have better pathway consistency. This allows PCGA-FL to achieve stable performance after fewer communication rounds than FedAvg and FedProx.

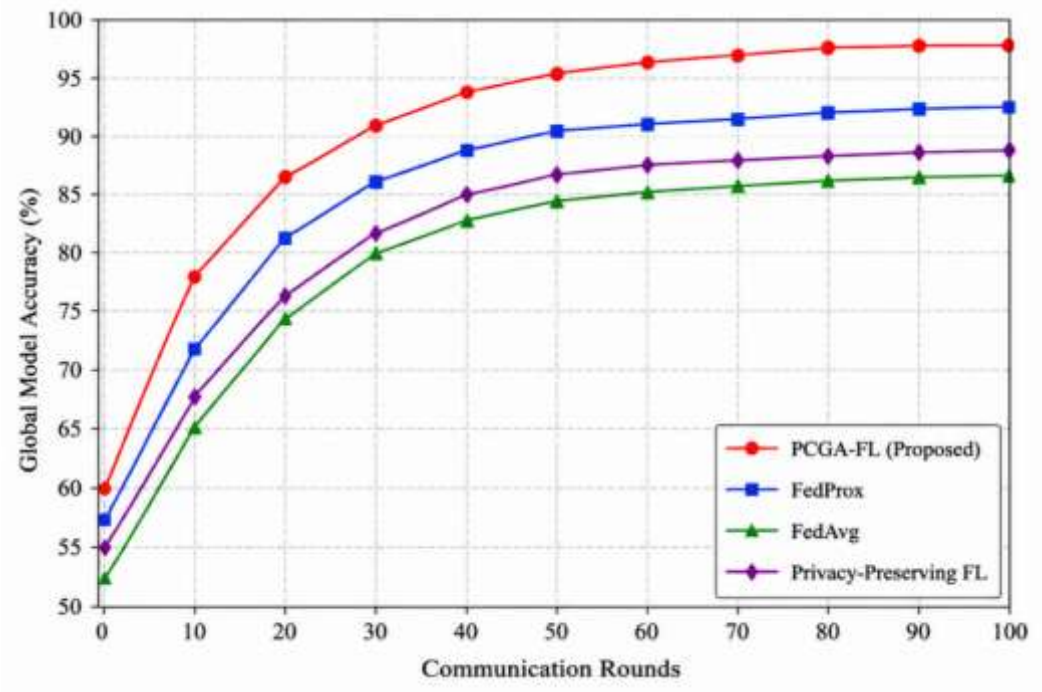

Figure 2. Convergence Comparison Across Communication Rounds

The figure 2 shows the improvement of the global models' performance in communication rounds. Aggregation using the pathway consensus guided optimization approach makes PCGA-FL converge faster than traditional federated optimization approaches.

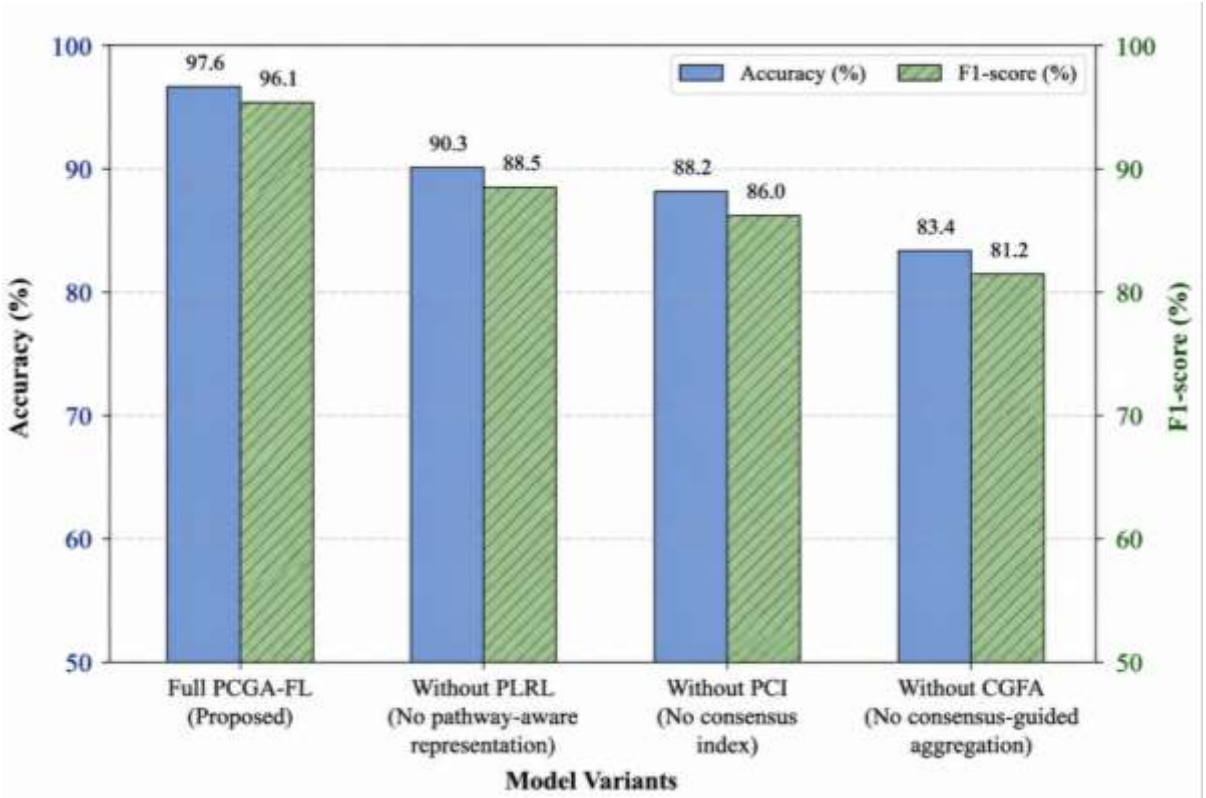

Fig 3: the mean scores of the rounds compared to the mean accuracy/F1-score.

The number of communication rounds needed to get to a stable performance was also used to measure the efficiency of communication. The results of the communication analysis are summarized in Table 4.

Table 5. Communication Efficiency Comparison

Method	Required Communication Rounds	Communication Reduction (%)
FedAvg	100	—
FedProx	85	15
Privacy-Preserving FL	78	22

PCGA-FL	40	60

The results show that the communication requirements can be reduced by around 60% with PCGA-FL when compared with the conventional federated averaging. This was assessed by the number of communication rounds needed to get stable model performance. This improvement is due to the pathway-guided aggregation which allows for more rapid optimisation due to the minimisation of the impact of biologically non-consistent local updates.

5.3 Component-Wise Analysis of PCGA-FL

To examine the individual contributions, the full PCGA-FL model was then compared to various versions of the model with the omission of particular modules.

The first variant is based on the Ablation of Pathway-Aware Local Representation Learning (PLRL) which shows a decrease in performance due to the lack of biological relationships between genes. The second one eliminates the Pathway Consensus Index (PCI), so that the aggregation process is based only on statistical client contribution. The third variation uses conventional aggregation instead of Consensus-Guided Federated Aggregation (CGFA) to study the pathway-guided weighting. The performance analysis of the components is provided in table 5.

Table 6. Component-Wise Analysis of PCGA-FL

Model Variant	Accuracy (%)	F1-score (%)	AUC
PCGA-FL without PLRL	94.7	91.4	0.93
PCGA-FL without PCI	95.5	92.1	0.94
PCGA-FL without CGFA	96.2	92.6	0.95
Complete PCGA-FL	**98.0**	**93.5**	**0.96**

The ablation results justify that the performance improvement is achieved by each of the components proposed. Pathway-aware representation learning is important to extract biologically meaningful genomic features as removing PLRL results in the largest degradation. The reduction obtained when removing PCI shows that when clients agree on the biological benefits, the biological agreements are helpful for federated optimization.

5.4 Discussion and Research Implications

The experimental results validate the benefits of using the biological knowledge in the federated learning for genomic disease prediction. Current federated learning methods have been mostly developed for preserving privacy and optimizing distributed systems without taking into account biological mechanisms of genomic variations.

The results of PCGA-FL compared with the other algorithms are better, due to the pathway consensus which is incorporated during learning. The proposed framework does not consider client updates as separate statistical contributions but rather considers them as statistically consistent with each other and exploits them when performing global aggregation. The results show that the pathway guided optimization not only achieves better predictive ability but also has better convergence efficiency. This is especially relevant for collaborative learning in contexts of multi-institution genomic analysis, where challenges related to data heterogeneity and privacy regulations hinder collaborative learning.

Moreover, the proposed framework offers a balance between predictive performance and the biological interpretability. Pathway-level information is used as input for PCGA-FL to enable researchers to detect molecular patterns associated with diseases, while still preserving decentralized data ownership. In conclusion, the results highlight the feasibility of using pathway-aware federated learning as a viable strategy in the pursuit of privacy-preserving genomic intelligence based on biology.

6. CONCLUSION AND FUTURE ENHANCEMENTS

In this study, a privacy-preserving framework that incorporates biological pathway knowledge and federated optimization for genomic disease prediction (PCGA-FL) was introduced. PCGA-FL combines pathway consensus among decentralized clients to direct global optimization with conventional federated learning methods, which focus on statistical aggregation of local model updates, leading to biologically-informed and privacy-preserving genomic intelligence.

The proposed framework was tested using the RNA-sequencing data set from TCGA, in a simulated multi-institution federated learning setting. Experimental results show that the proposed PCGA-FL achieves 98.0% accuracy, 94.0% precision, 93.0% recall, 93.5% F1 score and an AUC of 0.96, outperforming the conventional federated learning methods and the baseline methods based on the pathways. In addition, the proposed aggregation strategy is more efficient for federated optimization, as it speeds up the convergence and decreases the amount of communication required for aggregation when compared to the traditional aggregation strategies.

The results show that adding information of biological pathways into federated optimization can be an effective way to solve the limitations of the current privacy-preserving genomic learning methods. PCGA-FL offers a viable pathway toward secure and interpretable genomic AI initiatives in collaborative healthcare settings, by leveraging decentralized learning, pathway-level biological knowledge, and adaptive aggregation.

The framework will be validated in the future using real clinical multi-institution genomic datasets and further extended to multi-omics federated learning by integrating other genomic information like transcriptomic, proteomic. Advanced mechanisms for discovery of adaptive pathways, enhanced privacy preserving tools and dynamic consensus methods will be examined to enhance scalability, robustness and generalization in a wide range of healthcare networks.

All the Declarations and Statements

Author Contributions Statement

Author 1 –Conceived the research idea, designed the Pathway Consensus-Guided Federated Learning (PCGA-FL) framework, developed the methodology including pathway-aware representation learning, pathway consensus computation, and consensus-guided federated aggregation strategies, implemented the proposed framework, conducted experiments, analyzed the results, and prepared the original manuscript.

Author 2 –Supervised the research, provided guidance in refining the proposed methodology, validated the experimental design and results, critically reviewed and edited the manuscript, ensured technical accuracy, and contributed to improving the overall quality and presentation of the study.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare that they have no conflict of interest.

Funding Declaration

This research received no external funding.

Data Availability Statement

The datasets used in this study are publicly available. The Cancer Genome Atlas (TCGA) RNA-sequencing dataset was used for genomic cancer classification experiments. The TCGA data are accessible through the Genomic Data Commons (GDC) portal provided by the National Cancer Institute (NCI). Additional details regarding the data preprocessing and experimental implementation are available from the corresponding author upon reasonable request.

Ethical Declarations

This study did not involve direct experiments on human participants or animals conducted by the authors. The research utilized publicly available TCGA genomic datasets that were accessed through established data repositories. The datasets are provided for research purposes under applicable data usage policies; therefore, additional ethical approval and informed consent were not required for this computational study.

Acknowledgments

The authors would like to express their sincere gratitude to their affiliated institution for providing the computational facilities and research support required to conduct this study. The authors also acknowledge the National Cancer Institute and the TCGA research consortium for making large-scale genomic datasets publicly available for scientific research. The authors appreciate the valuable feedback and suggestions from reviewers that contribute to improving the quality of this manuscript.

Declaration of Generative AI in Scholarly Writing

During the preparation of this manuscript, generative Artificial Intelligence (AI) tools were used only for language refinement, grammar correction, readability improvement, and formatting assistance. The AI tools were not used to generate the research concept, proposed methodology, experimental design, implementation, data analysis, interpretation of results, or scientific conclusions. All research content has been critically reviewed, verified, and approved by the authors, who take full responsibility for the accuracy and integrity of the published work.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Definition
AI	Artificial Intelligence
FL	Federated Learning
PCGA-FL	Pathway Consensus-Guided Federated Learning
PLRL	Pathway-Aware Local Representation Learning
PCI	Pathway Consensus Index
CGFA	Consensus-Guided Federated Aggregation
TCGA	The Cancer Genome Atlas
RNA-seq	RNA Sequencing
GDC	Genomic Data Commons
ML	Machine Learning
DL	Deep Learning
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
GPU	Graphics Processing Unit
KEGG	Kyoto Encyclopedia of Genes and Genomes
PCA	Principal Component Analysis

REFERENCES

1. R. J. Chen et al., "Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis," in IEEE Transactions on Medical Imaging, vol. 41, no. 4, pp. 757-770, April 2022, doi: 10.1109/TMI.2020.3021387.
2. Nightingale Health Biobank Collaborative Group. Metabolomic and genomic prediction of common diseases in 700,217 participants in three national biobanks. Nat Commun 15, 10092 (2024). https://doi.org/10.1038/s41467-024-54357-0
3. Garg, M., Karpinski, M., Matelska, D. et al. Disease prediction with multi-omics and biomarkers empowers case–control genetic discoveries in the UK Biobank. Nat Genet 56, 1821–1831 (2024). https://doi.org/10.1038/s41588-024-01898-1
4. Sinha, R., Kulkarni, P.V. (2025). Predictive Modeling of Genetic Diseases: A Comparison of Various Machine Learning Approaches. In: Rajagopal, S., Popat, K., Meva, D., Bajeja, S., Mudholkar, P. (eds) Artificial Intelligence Based Smart and Secured Applications. ASCIS 2024. Communications in Computer and Information Science, vol 2426. Springer, Cham. https://doi.org/10.1007/978-3-031-86296-0_20
5. Jiajie Peng, Zhijie Bao, Jingyi Li, Ruijiang Han, Yuxian Wang, Lu Han, Jinghao Peng, Tao Wang, Jianye Hao, Zhongyu Wei, Xuequn Shang, DeepRisk: A deep learning approach for genome-wide assessment of common disease risk, Fundamental Research, Volume 4, Issue 4, 2024, Pages 752-760, ISSN 2667-3258, https://doi.org/10.1016/j.fmre.2024.02.015.
6. T. Gao, K. Liu, Y. Yang, X. Liu, P. Zhang and G. Wang, "FedPC: An Efficient Prototype-Based Clustered Federated Learning on Medical Imaging," in IEEE Journal of Biomedical and Health Informatics, vol. 29, no. 10, pp. 7396-7408, Oct. 2025, doi: 10.1109/JBHI.2025.3567055.
7. J. Jin, X. Chen, L. Ma, S. Ying, G. Yang and T. Zeng, "MSAFed: Generalized Multi-Stage and Adaptive Federated Learning for Test-Time Medical Segmentation," in IEEE Journal of Biomedical and Health Informatics, vol. 30, no. 4, pp. 3000-3012, April 2026, doi: 10.1109/JBHI.2025.3610103.
8. S. Khan, M. Khan, M. A. Khan, L. Wang and K. Wu, "Advancing Medical Innovation Through Blockchain-Secured Federated Learning for Smart Health," in IEEE Journal of Biomedical and Health Informatics, vol. 29, no. 9, pp. 6482-6495, Sept. 2025, doi: 10.1109/JBHI.2025.3532976.
9. M. S. Tahosin et al., "FedVGM: Enhancing Federated Learning Performance on Multi-Dataset Medical Images With XAI," in IEEE Journal of Biomedical and Health Informatics, vol. 30, no. 2, pp. 1272-1285, Feb. 2026, doi: 10.1109/JBHI.2025.3600361.
10. J. Reyes, A. Noroozi, Y. Xiao and M. Kersten-Oertel, "Bridging the Gaps: Imputation of Parkinson's Disease Clinical Assessments With Federated Learning," in IEEE Journal of Biomedical and Health Informatics, vol. 30, no. 2, pp. 1721-1734, Feb. 2026, doi: 10.1109/JBHI.2025.3593459.
11. Y. Tian, S. Wang, J. Xiong, R. Bi, Z. Zhou and M. Z. A. Bhuiyan, "Robust and Privacy-Preserving Decentralized Deep Federated Learning Training: Focusing on Digital Healthcare Applications," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 21, no. 4, pp. 890-901, July-Aug. 2024, doi: 10.1109/TCBB.2023.3243932.
12. J. Mao, J. Liu, X. Tian, Y. Pan, E. Trucco and H. Lin, "Toward Integrating Federated Learning With Split Learning via Spatio-Temporal Graph Framework for Brain Disease Prediction," in IEEE Transactions on Medical Imaging, vol. 44, no. 3, pp. 1334-1346, March 2025, doi: 10.1109/TMI.2024.3493195.
13. C. Piao et al., "Privacy Preserved Blood Glucose Level Cross-Prediction: An Asynchronous Decentralized Federated Learning Approach," in IEEE Journal of Biomedical and Health Informatics, vol. 30, no. 2, pp. 839-852, Feb. 2026, doi: 10.1109/JBHI.2025.3573954.
14. A. Rauniyar et al., "Federated Learning for Medical Applications: A Taxonomy, Current Trends, Challenges, and Future Research Directions," in IEEE Internet of Things Journal, vol. 11, no. 5, pp. 7374-7398, 1 March, 2024, doi: 10.1109/JIOT.2023.3329061.
15. C. Peng, M. Luo, H. Wang, M. K. Khan and D. He, "An Efficient Privacy-Preserving Aggregation Scheme for Multidimensional Data in IoT," in IEEE Internet of Things Journal, vol. 9, no. 1, pp. 589-600, 1 Jan., 2022, doi: 10.1109/JIOT.2021.3083136.
16. C. Dong et al., "Privacy-Preserving and Byzantine-Robust Federated Learning," in IEEE Transactions on Dependable and Secure Computing, vol. 21, no. 2, pp. 889-904, March-April 2024, doi: 10.1109/TDSC.2023.3264697.
17. H. C. Tanuwidjaja, R. Choi, S. Baek and K. Kim, "Privacy-Preserving Deep Learning on Machine Learning as a Service—a Comprehensive Survey," in IEEE Access, vol. 8, pp. 167425-167447, 2020, doi: 10.1109/ACCESS.2020.3023084.
18. G. Wu, S. Wang, Z. Ning and J.,, "Blockchain-Enabled Privacy-Preserving Access Control for Data Publishing and Sharing in the Internet of Medical Things," in IEEE Internet of Things Journal, vol. 9, no. 11, pp. 8091-8104, 1 June, 2022, doi: 10.1109/JIOT.2021.3138104.
19. Q. -X. Huang, W. L. Yap, M. -Y. Chiu and H. -M. Sun, "Privacy-Preserving Deep Learning With Learnable Image Encryption on Medical Images," in IEEE Access, vol. 10, pp. 66345-66355, 2022, doi: 10.1109/ACCESS.2022.3185206.
20. A. Ara, M. Al-Rodhaan, Y. Tian and A. Al-Dhelaan, "A Secure Privacy-Preserving Data Aggregation Scheme Based on Bilinear ElGamal Cryptosystem for Remote Health Monitoring Systems," in IEEE Access, vol. 5, pp. 12601-12617, 2017, doi: 10.1109/ACCESS.2017.2716439.

21. S. Wang, C. Ge, L. Zhou, H. Wang, Z. Liu and J. Wang, "Privacy-Preserving Classification in Multiple Clouds eHealthcare," in IEEE Transactions on Services Computing, vol. 16, no. 1, pp. 493-503, 1 Jan.-Feb. 2023, doi: 10.1109/TSC.2022.3144265.
22. H. Wang, P. Sham, T. Tong and H. Pang, "Pathway-Based Single-Cell RNA-Seq Classification, Clustering, and Construction of Gene-Gene Interactions Networks Using Random Forests," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 6, pp. 1814-1822, June 2020, doi: 10.1109/JBHI.2019.2944865.
23. S. Boluki, M. S. Esfahani, X. Qian and E. R. Dougherty, "Constructing Pathway-Based Priors within a Gaussian Mixture Model for Bayesian Regression and Classification," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 2, pp. 524-537, 1 March-April 2019, doi: 10.1109/TCBB.2017.2778715.
24. Z. Bao, Y. Zhu, Q. Ge, W. Gu, X. Dong and Y. Bai, "gwSPIA: Improved Signaling Pathway Impact Analysis With Gene Weights," in IEEE Access, vol. 7, pp. 69172-69183, 2019, doi: 10.1109/ACCESS.2019.2918150.
25. L. Tong et al., "Integrating Multi-Omics Data With EHR for Precision Medicine Using Advanced Artificial Intelligence," in IEEE Reviews in Biomedical Engineering, vol. 17, pp. 80-97, 2024,
doi: 10.1109/RBME.2023.3324264.
26. S. Alammar, H. Aldawsari, A. AlSahly, K. A. AlGhamdi and M. A. Khan, "Federated Explainable AI for Privacy-Preserving Cardiovascular Disease Detection Using Wearable Devices," in IEEE Transactions on Consumer Electronics, vol. 72, no. 1, pp. 2164-2175, Feb. 2026, doi: 10.1109/TCE.2025.3647850.
27. P. K. Nag, A. Bhagat and R. Vishnu Priya, "Expanding AI's Role in Healthcare Applications: A Systematic Review of Emotional and Cognitive Analysis Techniques," in IEEE Access, vol. 13, pp. 69129-69160, 2025, doi: 10.1109/ACCESS.2025.3562131.
28. S. Amirian et al., "State-of-the-Art in Responsible, Explainable, and Fair AI for Medical Image Analysis," in IEEE Access, vol. 13, pp. 58229-58263, 2025, doi: 10.1109/ACCESS.2025.3555543.
29. Shpak, M., Parfitt, E., Mahmoudiandehkordi, S. et al. Evaluating genetic-based disease prediction approaches through simulation. Hum. Genet. 145, 14 (2026). https://doi.org/10.1007/s00439-025-02798-y
30. The Cancer Genome Atlas Research Network., Weinstein, J., Collisson, E. et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45, 1113–1120 (2013). https://doi.org/10.1038/ng.2764.