Toward Reproducible and Privacy-Preserving AI Systems for Multi-Modal Biomarker Discovery

Victor Ikenna Kanu, Simeon Okechukwu Ajakwe, Dong-Seong Kim
Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi, South Korea
(kanuxavier, simeonajlove)@gmail.com, dskim@kumoh.ac.kr

***Abstract*—Artificial intelligence (AI)–driven healthcare analytics face a critical reproducibility challenge, with many models failing to generalize across heterogeneous and distributed datasets. This survey reviews the evolution of computational frameworks for biomarker discovery, from early correlation-based networks to causal Bayesian inference and deep multimodal learning architectures. We identify limited auditability, interpretability, and deterministic execution as primary ICT engineering bottlenecks. To address these challenges, we outline a next-generation AI systems framework leveraging Federated Learning and Explainable AI (XAI) to enable reproducible, privacy-preserving, and trustworthy analytics across decentralized institutions.**

***Index Terms*—Neurodegenerative Diseases, Multi-omics Integration, Reproducible AI, Federated Learning, Explainable AI (XAI).**

I. Introduction

The rapid proliferation of high-throughput omics technologies has transformed the landscape of neurodegenerative disease research, enabling the identification of complex molecular signatures associated with Alzheimer's Disease (AD) and Parkinson's Disease (PD). Traditional single-modality approaches often fail to capture the heterogeneous nature of these pathologies, leading to a significant "mRNA-protein gap" where transcriptomic changes do not reliably predict downstream protein-level dysfunction. Consequently, the field is shifting toward AI-driven multi-omics integration, which fuses genomics, proteomics, and metabolomics to uncover robust biomarkers for early diagnosis and precision medicine [1]. However, despite these advancements, the clinical translation of AI models remains hindered by a critical lack of reproducibility; a recent meta-analysis revealed that only 21% of health AI studies release code and fewer than 23% validate findings on external datasets, resulting in models that fail to generalize across diverse patient cohorts [2]. This survey reviews the evolution of computational frameworks from correlation-based networks to causal inference and explainable AI, highlighting the urgent engineering need for deterministic, auditable pipelines to bridge the gap between discovery and clinical utility.

II. AI-Driven Biomarker Discovery in Neurodegeneration

The application of AI to neurodegeneration has evolved from descriptive correlation to predictive causal inference and deep phenotyping. Early computational frameworks utilized Weighted Gene Co-expression Network Analysis (WGCNA) to identify modules of co-regulated genes associated with regional vulnerability in Alzheimer's Disease (AD), establishing spatial maps of pathology [3]. However, correlation does not imply causation. To address this, authors [4] introduced causal Bayesian networks anchored by cis-regulatory genetic variants (cis-eSNPs), successfully identifying TYROBP as a master regulator of the immune-microglia module—a finding validated by in vitro perturbation.

Recent frameworks have expanded this approach through multi-omics integration to resolve data heterogeneity [5]. Authors [6] demonstrated that Commonality Analysis across transcriptomics, proteomics, and metabolomics could identify conserved metabolic dysfunctions (specifically Vitamin B pathways) in both human and mouse models, offering robust cross-species biomarkers. Most recently, the field has adopted Transformer architectures to handle high-dimensional multimodal data. Authors [7] developed a unified Vision-Text Transformer that achieved 98.75% diagnostic accuracy for AD. Crucially, this study integrated Explainable AI (XAI) via LIME to visualize decision boundaries, confirming that the model's predictions were driven by biologically valid features, such as hippocampal atrophy, rather than statistical artifacts. A comparative summary of these computational frameworks, highlighting their contributions to reproducibility and specific limitations, is provided in Table I.

TABLE I
COMPARATIVE ANALYSIS OF COMPUTATIONAL FRAMEWORKS FOR NEURODEGENERATIVE BIOMARKER DISCOVERY

Generation	**Methodology**	**Study**	**Reproducibility Contribution**	**Limitations**
Gen 1: Correlation	WGCNA (Co-expression Networks)	[3]	Mapped spatial vulnerability in AD brains.	Identifies correlation only; lacks causal direction.
Gen 2: Causal	Bayesian Networks (Genetic Anchors)	[4]	Identified causal drivers (e.g., TYROBP) using cis-eSNPs.	Computationally expensive; requires large sample sizes.
Gen 2: Multi-Layer	Commonality Analysis (Venn Integration)	[6]	Validated metabolic pathways across species (Human/Mouse).	Intersection methods discard unique, modality-specific data.
Gen 3: Deep	Deep Phenotyping (18-Platform Integration)	[8]	Established ”Mutual Best Hits” for cross-platform validation.	High cost/complexity to generate 18-layer datasets.
Gen 3: Explainable	Transformers + XAI (LIME)	[7]	Validated ”Black Box” decisions against clinical biomarkers.	Vision Transformers require massive data to outperform CNNs.

III. Reproducibility Challenges in Biomarker Discovery

Despite the promise of AI in neurodegeneration, the field faces a systemic "reproducibility crisis" that limits clinical translation. A comprehensive meta-analysis of 511 Machine Learning for Health (MLH) studies identified a substantial lack of transparency: only 21% of papers released their code, and fewer than 55% utilized public datasets. Notably, only 23% of studies assessed their models on external datasets from multiple institutions. This deficiency in conceptual replicability, defined as the ability of a model to perform effectively in new clinical environments, results in widespread dataset overfitting, where models capture site-specific artifacts rather than generalizable biological signals [2].

Beyond data availability, the software engineering architecture underlying validation frameworks remains underdeveloped. A meta-analysis of single-cell benchmarks found that although code sharing is increasing, fewer than 8% of studies use containerization technologies such as Docker to ensure reproducible computational environments, limiting the extensibility of many pipelines. Moreover, the emerging field of digital biomarkers lacks standardized validation criteria, leading to fragmented datasets that cannot be readily pooled for robust model training [9]. Unless these foundational engineering challenges are addressed, AI-driven biomarkers are unlikely to progress beyond academic research and become reliable clinical tools.

IV. TOWARD DETERMINISTIC AND AUDITABLE AI PIPELINES

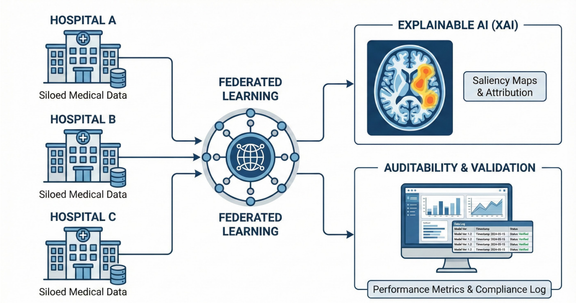

Fig. 1. Proposed Federated Learning framework integrating Explainable AI (XAI) for reproducible biomarker discovery.

To address the previously identified engineering gaps, future biomarker discovery frameworks should emphasize accessibility and auditability rather than focusing solely on algorithmic complexity. Recent developments indicate that democratizing access to multi-omics networks is essential for effective validation. Authors [8] introduced the ”Molecular Human” framework, which integrates 18 high-throughput platforms into an open-access web server (Comics). This tool enables users without coding expertise to query molecular interactions in real time, thereby shifting the paradigm from static code repositories to dynamic and auditable exploration. Similarly, authors [10] developed ILINCS, a platform that precomputes over one billion connections between transcriptomic and proteomic signatures. This system supports deterministic “Connectivity Analysis,” allowing researchers to consistently identify therapeutic targets, such as mTOR signaling, that are often overlooked by traditional stochastic pathway analyses.

Nevertheless, centralized web tools are insufficient to address the privacy constraints inherent in clinical data. Achieving conceptual replicability without compromising patient privacy requires adopting the Federated Learning (FL) architecture illustrated in Fig 1. FL enables the training of complex models, such as the Vision Transformers described by authors [7], across decentralized datasets from multiple institutions, including hospitals, without sharing raw patient data. By separating model training from data centralization, FL provides a deterministic engineering approach for validating biomarkers across diverse populations, thereby ensuring that AI-driven discoveries remain robust, unbiased, and clinically reliable.

V. CONCLUSION

The evolution of biomarker discovery from descriptive correlation-based networks to causal and deep phenotyping frameworks represents a significant maturation in neurodegenerative research . However, the clinical utility of these models remains stalled by a fundamental reproducibility crisis, where ”black box” AI trained on siloed datasets fails to generalize . To bridge this gap, future ICT convergence research must prioritize auditability and interpretability as core engineering requirements. Specifically, we identify the adoption of Federated Learning (FL) architectures as the critical next step; FL enables the training of complex, data-hungry models across multi-institutional cohorts without compromising patient privacy, ensuring that AI-driven biomarkers are robust, unbiased, and clinically trustworthy .

ACKNOWLEDGMENT

This work was partly supported by Innovative Human Resource Development for Local Intellectualization program through the IITP grant funded by the Korea government (MSIT) (IITP-2025-RS-2020-II201612, 33%) and by Priority Research Centers Program through the NRF funded by the MEST (2018R1A6A1A03024003, 33%) and by the MSIT, Korea, under the ITRC support program (IITP-2025-RS-2024-00438430, 34%)).

REFERENCES

[1] Z. Jiang, H. Zhang, Y. Gao, and Y. Sun, “Multi-omics strategies for biomarker discovery and application in personalized oncology,” *Molecular Biomedicine*, vol. 6, no. 1, p. 115, 2025.

[2] M. B. McDermott, S. Wang, N. Marinsek, R. Ranganath, L. Foschini, and M. Ghassemi, “Reproducibility in machine learning for health research: Still a ways to go,” *Science translational medicine*, vol. 13, no. 586, p. eabb1655, 2021.

[3] M. Wang, P. Roussos, A. McKenzie, X. Zhou, Y. Kajiwara, K. J. Brennand, G. C. De Luca, J. F. Crary, P. Casaccia, J. D. Buxbaum *et al.*, “Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to alzheimer’s disease,” *Genome medicine*, vol. 8, no. 1, p. 104, 2016.

[4] B. Zhang, C. Gaiteri, L.-G. Bodea, Z. Wang, J. McElwee, A. A. Podtelezhnikov, C. Zhang, T. Xie, L. Tran, R. Dobrin *et al.*, “Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer’s disease,” *Cell*, vol. 153, no. 3, pp. 707–720, 2013.

[5] V. I. Kanu, J. Isong, S. O. Ajakwe, T. Jun, and D.-S. Kim, “Blockchain-enabled framework for efficient and interoperable proteomic data management,” in *2025 Sixteenth International Conference on Ubiquitous and Future Networks (ICUFN)*. IEEE, 2025, pp. 666–671.

[6] P. Kodam, R. Sai Swaroop, S. S. Pradhan, V. Sivaramakrishnan, and R. Vadrevu, “Integrated multi-omics analysis of alzheimer’s disease shows molecular signatures associated with disease progression and potential therapeutic targets,” *Scientific reports*, vol. 13, no. 1, p. 3695, 2023.

[7] H. Anzum, N. S. Sammo, and S. Akhter, “Leveraging transformers and explainable ai for alzheimer’s disease interpretability,” *PLoS One*, vol. 20, no. 5, p. e0322607, 2025.

[8] A. Halama, S. Zaghlool, G. Thareja, S. Kader, W. Al Muftah, M. Mook-Kanamori, H. Sarwath, Y. A. Mohamoud, N. Stephan, S. Ameling *et al.*, “A roadmap to the molecular human linking multiomics with population traits and diabetes subtypes,” *Nature communications*, vol. 15, no. 1, p. 7111, 2024.

[9] V. I. Kanu, S. O. Ajakwe, J. M. Lee, and D.-S. Kim, “Deterministic protein structure and binding site analysis through blockchain-integrated workflow verification,” *ICT Express*, 2025.

[10] M. Pilarczyk, M. Fazel-Najafabadi, M. Kouril, B. Shamsaei, J. Vasiliauskas, W. Niu, N. Mahi, L. Zhang, N. A. Clark, Y. Ren *et al.*, “Connecting omics signatures and revealing biological mechanisms with ilincs,” *Nature communications*, vol. 13, no. 1, p. 4678, 2022.