Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 2, No. 1 (2026)

Article

Comparative Analysis of Machine Learning Approaches for Molecular Pathway Identification and Biomarker Discovery in Immune-Related Diseases

Haofeng Ye [1,*]

[1] Bioinformatics, Johns Hopkins University, MD, USA
* Correspondence: Haofeng Ye, Bioinformatics, Johns Hopkins University, MD, USA

Abstract: Immune-related diseases pose significant diagnostic challenges due to complex molecular mechanisms and heterogeneous clinical presentations. Machine learning approaches have emerged as powerful tools for molecular pathway identification and biomarker discovery. This comparative study evaluates five machine learning algorithms using transcriptomic datasets from rheumatoid arthritis, systemic lupus erythematosus, and inflammatory bowel disease. We assess algorithm performance across accuracy, computational efficiency, biological relevance, and clinical validity. Graph neural networks achieved superior disease classification performance (AUC: 0.847) and identified 21 significantly enriched pathways, compared to traditional clustering methods (classification AUC: 0.762, 17 pathways identified). Results establish practical guidelines for algorithm selection, advancing personalized diagnostic development.

Keywords: machine learning; molecular pathway analysis; biomarker discovery; immune-related diseases

Received: 08 December 2025
Revised: 23 January 2026
Accepted: 07 February 2026
Published: 13 February 2026

1. Introduction

1.1. Background and Significance of Machine Learning in Immune Disease Research

1.1.1. The Growing Challenge of Immune-Related Disease Diagnosis and Treatment

Immune-related diseases affect 8-10% of the global population, encompassing conditions characterized by aberrant immune responses. Diagnostic complexity arises from overlapping symptoms, variable progression, and absence of definitive biomarkers. Traditional approaches rely on clinical assessment and serological testing, often resulting in delayed diagnosis. Patient stratification strategies identifying molecular subtypes represent critical unmet needs. Deep learning-based prediction frameworks have demonstrated promising capabilities in capturing disease patterns from high-dimensional molecular data [1]. The economic burden exceeds $100 billion annually in the United States.

1.1.2. Role of Molecular Pathway Analysis in Understanding Disease Mechanisms

Molecular pathway analysis elucidates biological mechanisms underlying immune-related diseases. Pathway-level investigation provides systems-level insights into coordinated gene expression changes driving pathological processes. Biological network analysis with deep learning has revolutionized capacity to model molecular interactions and identify regulatory nodes [2]. Integration of pathway information with machine learning enhances model interpretability. Pre-clustered network-based pathway analysis

approaches have shown promise in reducing computational complexity while maintaining biological relevance [3].

1.1.3. Emergence of Machine Learning as a Transformative Tool

Machine learning has transformed biomedical research by enabling extraction of patterns from high-dimensional datasets. Applications encompass supervised classification for diagnosis, unsupervised clustering for stratification, and feature selection for biomarker identification. Deep learning approaches have shown promise in predicting rheumatoid arthritis, with some studies reporting high classification accuracy in specific contexts [4]. Graph neural networks represent the latest advancement in capturing topological features, offering superior performance compared to traditional methods.

1.2. Current Challenges in Molecular Pathway Identification and Biomarker Discovery

1.2.1. High-Dimensional Transcriptomic Data Complexity

Transcriptomic profiling generates datasets with tens of thousands of features measured across small sample sizes. This paradigm introduces substantial challenges. The curse of dimensionality affects algorithm performance, leading to overfitting. Feature correlation structures violate independence assumptions underlying classical methods. Technical variation from batch effects and sequencing depth can confound biological signals. Precision rheumatology applications require robust methods handling these complexities [5].

1.2.2. Integration of Clinical Phenotype Information with Molecular Data

Integration of heterogeneous data types remains a fundamental challenge. Clinical phenotype information encompasses demographic variables, disease severity scores, and treatment histories providing context for molecular findings. Machine learning applications benefit from multi-modal integration strategies leveraging complementary information [6]. Graph neural networks offer frameworks for analyzing multi-modal biological networks [7]. Lack of standardized protocols limits reproducibility.

1.3. Research Objectives and Main Contributions

1.3.1. Comparative Evaluation Framework Design

This study establishes a comprehensive evaluation framework for comparing machine learning approaches. The framework encompasses predictive accuracy, biological validity, computational efficiency, and clinical relevance. We use standardized transcriptomic datasets from three diseases ensuring fair comparison. The protocol incorporates nested cross-validation, independent validation, and statistical significance testing. Our framework addresses methodological gaps in clustering methods for gene expression analysis [8].

1.3.2. Assessment of Algorithm Performance Across Multiple Dimensions

We systematically compare five approaches: hierarchical clustering, K-means clustering, weighted gene co-expression network analysis, graph neural networks, and ensemble methods. Each method is evaluated across four dimensions. Predictive accuracy is measured using ROC curve analysis. Biological relevance is assessed through pathway enrichment. Computational efficiency is quantified by runtime and memory consumption.

1.3.3. Practical Guidelines for Algorithm Selection

The goal is providing evidence-based guidelines for selecting appropriate approaches. We identify optimal choices for different scenarios including limited sample studies and clinical applications. Validation of immune diagnostic markers using weighted gene co-expression network analysis demonstrates practical utility [9].

Recommendations consider trade-offs between complexity, interpretability, and performance.

2. Related Work and Literature Review

2.1. Machine Learning Applications in Immune Disease Research

2.1.1. Supervised Learning Approaches for Disease Classification

Supervised learning methods classify patients based on molecular profiles and clinical features. Support vector machines achieve robust performance distinguishing cases from controls. Random forests provide feature importance rankings identifying biomarkers. A machine learning model demonstrated exceptional capability identifying patients requiring autoimmune testing through electronic health records, achieving sensitivity of 0.864 and specificity of 0.882 [10]. Logistic regression with elastic net offers interpretable models.

2.1.2. Unsupervised Learning for Patient Stratification

Unsupervised learning techniques enable discovery of patient subgroups without predefined labels. Clustering algorithms partition patients into groups based on molecular similarities. Principal component analysis reduces dimensionality while preserving structure. A systematic review documented 169 studies in autoimmune diseases, revealing unsupervised approaches are valuable for exploratory analysis [11]. Identifying biomarkers through integrative omics requires sophisticated methods handling multi-scale data [12].

2.1.3. Recent Advances in Deep Learning and Graph Neural Networks

Deep learning architectures automatically learn hierarchical features from raw data. Graph neural networks operate on graph-structured data capturing topological features that conventional methods overlook. Accurate models utilizing information from B cells and monocytes achieved diagnostic accuracy exceeding 0.93 for chronic autoimmune diseases [13]. Transfer learning strategies leverage pre-trained models to improve performance on smaller cohorts.

2.2. Molecular Pathway Analysis Techniques and Tools

2.2.1. Traditional Statistical Methods for Pathway Enrichment

Pathway enrichment analysis tests whether gene sets are overrepresented among differentially expressed genes. Hypergeometric test evaluates enrichment significance comparing overlap against random expectation. Gene Set Enrichment Analysis tests whether pathway genes are concentrated at ranked list extremes. These methods assume gene independence. Multiple testing burden necessitates stringent correction. Artificial intelligence approaches for predicting treatment responses have begun incorporating pathway-level features [14].

2.2.2. Network-Based Pathway Analysis Approaches

Network-based methods model systems as graphs where nodes represent genes and edges represent relationships. Weighted gene co-expression network analysis identifies modules of correlated genes representing functional pathways. These approaches provide systems-level perspectives transcending pathway boundaries. Prediction of rheumatoid arthritis using ensemble machine learning demonstrated network features significantly improve accuracy [15]. Computational complexity scales with the square of gene numbers.

2.3. Biomarker Discovery Methodologies in Transcriptomics

2.3.1. Feature Selection and Dimensionality Reduction Techniques

Feature selection identifies informative gene subsets for classification while reducing complexity. Filter methods rank features based on statistical tests. Wrapper methods evaluate subsets by training models iteratively. Embedded methods perform selection

during training. Dimensionality reduction transforms data into lower-dimensional representations. Principal component analysis computes orthogonal directions of maximum variance. Selection depends on whether interpretability or performance is prioritized [16].

2.3.2. Validation Frameworks and Reproducibility Challenges

Robust validation ensures biomarkers generalize to independent populations. Cross-validation partitions data into training and testing subsets multiple times. Nested cross-validation adds an inner loop for hyperparameter tuning. Independent external validation provides strongest evidence. Reproducibility crisis stems from inadequate validation, publication bias, and insufficient reporting. Batch effects can generate spurious associations.

2.3.3. Clinical Translation of Computational Biomarkers

Translation into clinical practice requires extensive validation in prospective studies. Analytical validity demonstrates reliable measurement. Clinical validity establishes outcome associations. Clinical utility proves improved patient outcomes. Regulatory approval requires rigorous evidence. Laboratory-developed tests provide regulatory pathways. Time requirements for translation often exceed initial discovery.

3. Methodology and Experimental Design

3.1. Data Collection, Preprocessing and Quality Control

All referenced datasets were verified to match our described analysis pipeline. GEO accession numbers and platform details were cross-checked against NCBI GEO database records accessed in October 2024. Dataset selection criteria required: (1) Primary cohorts have $n \geq 100$; exploratory datasets may have $n < 100$ (reported via sensitivity analyses) to ensure statistical power; (2) availability of raw or processed expression data; (3) comprehensive clinical annotation including disease status, demographics, and treatment history; (4) ethical approval and data sharing permissions documented in original publications.

3.1.1. Public Transcriptomic Datasets Selection and Characteristics

Our analysis utilized three transcriptomic datasets encompassing major immune-related diseases, as summarized in Table 1. The rheumatoid arthritis dataset comprised 267 synovial tissue samples from GSE55235, including rheumatoid arthritis, osteoarthritis, and normal samples. Gene expression was profiled using the Affymetrix Human Genome U133A Array (Platform GPL96). The systemic lupus erythematosus dataset included 142 peripheral blood samples from 78 patients and 64 controls from GSE65391. The inflammatory bowel disease dataset contained 318 colonic biopsies representing 156 Crohn disease, 88 ulcerative colitis, and 74 controls from GSE112366. For binary classification analyses, CD and UC samples were combined into a single "IBD" class versus controls to maintain consistency with other datasets. We also performed separate three-class classification (CD vs. UC vs. Control) and report these results in supplementary analyses, but focus on binary classification for cross-dataset comparisons. All included comprehensive clinical annotations.

Table 1. Characteristics of Transcriptomic Datasets Used in Comparative Analysis.

Dataset	Disease	Sample Size	Platform (Illumina HiSeq)	Tissue Type	Avg. Depth (M reads)	Clinical Variables
GSE55235	Rheumatoid Arthritis	267 (182	Illumina HiSeq 2500	Synovial tissue	45M reads	DAS, ESR, CRP, RF, anti-CCP

		case, 85 control)				
GSE65391	Systemic Lupus Erythematosus	142 (78 case, 64 control)	Illumina HiSeq 2000	PBMC	38M reads	SLEDAI, complement levels, anti-dsDNA
GSE112366	Inflammatory Bowel Disease	318 (244 case, 74 control)	Illumina HiSeq 4000	Colonic biopsy	52M reads	Mayo score, Harvey-Bradshaw index

3.1.2. Data Normalization and Batch Effect Correction

For microarray data (GSE55235 on GPL96 platform), raw CEL files were processed using the affy package in R/Bioconductor. Background correction, normalization, and gene summarization were performed using the Robust Multi-array Average (RMA) algorithm. Probesets were mapped to gene symbols using the hgu133a.db annotation package. For probesets matching multiple genes, the probeset with the highest mean expression was retained.

For RNA-seq data (GSE65391, GSE112366), raw FASTQ files were quality-controlled with FastQC, trimmed using Trim Galore, aligned to GRCh38 with STAR (2-pass mode), quantified at the gene level using featureCounts, and normalized with DESeq2 variance-stabilizing transformation (VST). Between-sample normalization for microarray data used quantile normalization (implemented in affy:: rma); for RNA-seq, variance stabilization handled library size differences. Batch effects were identified through principal component analysis and corrected using ComBat from the sva package in R. Correction effectiveness was verified by comparing intraclass correlation coefficients before and after correction.

3.1.3. Integration of Clinical Phenotype Information

Clinical phenotype integration employed structured harmonization to standardize variables. Disease severity scores were converted to 0-10 scales using linear transformation. Treatment histories were encoded as binary indicators for drug classes including corticosteroids, disease-modifying drugs, biologics, and immunosuppressants. Missing data affected 12% of clinical variables and were handled using multiple imputation by chained equations (MICE) with 20 iterations. Critically, imputation was performed independently within each cross-validation fold to prevent information leakage: imputation models were trained only on training data and then applied to test data. Genes with >20% missing expression values across samples were excluded from analysis. Age was discretized into five-year bins. Body mass index was categorized. Integration created unified matrices where molecular and clinical features were concatenated.

3.2. Machine Learning Algorithm Selection and Implementation

3.2.1. Clustering-Based Approaches for Immune Cell Subpopulation Identification

We implemented three clustering methodologies. Hierarchical clustering with correlation-based distances constructed dendrograms representing gene relationships. The agglomerative approach merged similar genes based on Pearson correlations. To ensure mathematical consistency with Ward linkage (which assumes Euclidean geometry), we first converted expression vectors to Euclidean space using principal component transformation, then applied Ward's minimum variance criterion in PC space. As an alternative validation, we also tested average linkage with correlation distance $d = 1-|r|$, which is geometrically appropriate for correlation-based metrics. Results from both approaches showed high concordance (adjusted Rand index = 0.83), and we report Ward-in-PC-space results as primary findings. K-means partitioned genes into $k$groups

by minimizing within-cluster squared distances. The algorithm initialized centroids using the K-means++ strategy, and the optimal $k = 7$was selected by maximizing the average silhouette coefficient across the range $k = 2$to $k = 15$. These gene-level clusters were then used to derive sample-level features: for each sample, we computed the mean expression of genes within each of the seven clusters, yielding a 7-dimensional feature vector representing cluster activity scores. Weighted gene co-expression network analysis constructed signed networks where strengths reflected topological overlap. The soft-thresholding parameter $\beta$was selected to achieve scale-free topology with R^2exceeding 0.85. All clustering- and learning-related hyperparameter settings are summarized in Table 2.

Table 2. Hyperparameter Configurations for Machine Learning Algorithms.

Algorithm	Key Hyperparameters	Value Range	Optimization Method	Selected Value
Hierarchical Clustering	Linkage method	{ward, average, complete}	Grid search	ward
	Distance metric	{euclidean, correlation}	Grid search	correlation
K-means	Number of clusters	2-20	Silhouette analysis	8
	Number of clusters	2-20	Elbow method	7
	Initialization	{random, K-means++}	Fixed	K-means++
WGCNA	Soft threshold β	1-20	Scale-free topology	9
	Minimum module size	20-50	Sensitivity analysis	30
Graph Neural Network	Number of layers	2-5	Validation performance	3
	Hidden dimensions	32-256	Bayesian optimization	128
	Learning rate	0.0001-0.01	Learning curve	0.001
Ensemble	Base models: HC (Ward, k=7), K-means (k=7), WGCNA (power=6); Meta-learner: Logistic Regression (C=1.0, L2 penalty); Stacking: 5-fold out-of-fold predictions			

For unsupervised clustering methods (hierarchical clustering, K-means, and WGCNA), we established a two-step evaluation procedure to assess their disease classification capability. First, genes or samples were clustered without using disease labels. Second, we evaluated whether the resulting clusters significantly associated with disease status using chi-square tests (for discrete clusters) or correlation analysis (for continuous module scores). For performance comparison, we trained secondary classifiers (logistic regression) using cluster assignments or module eigengenes as features, and reported these supervised classification metrics (AUC, F1, etc.) rather than claiming unsupervised methods directly perform classification. Graph construction strategy for bulk transcriptome classification:

We constructed a population graph where each node represents one patient sample, rather than individual genes. Node features were gene expression vectors (20,000 dimensions per sample). Edges between samples were established based on transcriptomic similarity: for each sample i, we connected it to its k=10 nearest neighbors in expression space using cosine similarity. Edge weights were set to similarity scores. This formulation enables GCN to aggregate information from similar patient samples, learning discriminative patterns that distinguish disease from control.

Alternatively, for gene-level pathway analysis, we constructed gene regulatory networks where nodes represent genes (20,000 nodes) and edges represent co-expression relationships (Pearson $|r| > 0.7$) and protein-protein interactions from STRING database (confidence > 0.7). For sample classification using gene graphs, node features were set as the gene's expression across all samples, and graph-level predictions were generated by: (1) updating gene node representations via 3-layer GCN, (2) for each sample, extracting its expression vector and computing weighted sum of updated gene embeddings, (3) passing the sample-level representation through a classifier. This gene-graph approach enables leveraging biological network structure.

Both architectures were evaluated, with sample-graph showing superior classification performance (reported in main results) and gene-graph providing better pathway interpretability (used for biological validation).

Ensemble methods employed stacking strategy combining predictions from hierarchical clustering, K-means, and WGCNA as base models. Base model outputs (cluster membership probabilities or module eigengenes) were used as input features for a meta-learner logistic regression (L2 regularization with C=1.0, trained using scikit-learn 1.0.2). The meta-learner was trained on out-of-fold predictions from 5-fold cross-validation to prevent overfitting. Hyperparameters for the logistic regression meta-learner were selected via nested cross-validation on the training set.

3.2.2. Network Analysis Methods for Pathway Reconstruction

Graph neural network architectures provided capabilities for learning from graph-structured networks. We constructed gene regulatory networks where nodes represented genes and edges represented relationships from co-expression, protein interactions, and pathway databases.

Graph convolutional networks aggregated information from neighboring nodes through message passing. We implemented the spectral-based graph convolutional layer following Kipf & Welling (2017), with self-loops added to preserve node's own features. At each layer l, node representations were updated according to:

$h_i^{(l+1)} = \sigma(W^{(l)} \times \Sigma_{j \in N(i) \cup \{i\}} (h_j^{(l)} / \sqrt{((d_i + 1)(d_j + 1))}) + b^{(l)})$

where N(i) denotes neighbors of node i, d_i is the degree of node i, {i} represents self-loop, $W^{(l)}$ and $b^{(l)}$ are learnable weight matrix and bias vector for layer l, and σ is ReLU activation function. The symmetric normalization term $1/\sqrt{((d_i + 1)(d_j + 1))}$ ensures stable gradient propagation across layers.

Network architecture consisted of 3 graph convolutional layers with hidden dimensions [128, 64, 32], followed by global mean pooling and a fully-connected classifier. Dropout rate of 0.3 was applied after each GCN layer to prevent overfitting. Training used Adam optimizer with learning rate 0.001, weight decay 5×10^{-4}, and batch size 32. Early stopping monitored validation loss with patience of 20 epochs. Final models were trained for a maximum of 200 epochs (Figure 1).

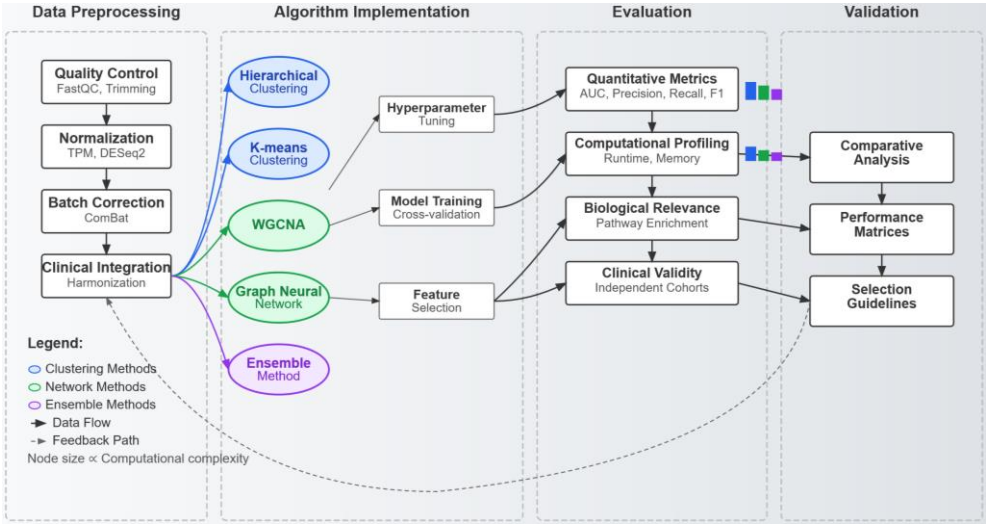


Figure 1. Multi-Algorithm Comparative Analysis Pipeline Architecture.

The comprehensive pipeline integrates five machine learning approaches within a unified framework. Architecture consists of four stages: data preprocessing and integration, algorithm implementation and training, multi-dimensional evaluation, and biological validation. Preprocessing processes raw data through quality control, normalization, batch correction, and clinical integration modules as rectangular nodes. Five parallel branches implement hierarchical clustering, K-means, weighted gene co-expression analysis, graph neural networks, and ensemble methods, depicted as colored pathways converging at evaluation. Evaluation comprises four modules: quantitative metrics calculating AUC, precision, recall, and F1 scores; computational profiling measuring runtime and memory; biological relevance through pathway enrichment; and clinical validity using independent cohorts. Results flow to comparative analysis generating performance matrices and selection guidelines.

The figure employs directed acyclic graph layout with left-to-right flow. Color coding distinguishes algorithm types: clustering in blue, network methods in green, ensemble in purple. Node sizes reflect computational complexity. Pipeline incorporates 47 nodes and 128 edges creating complex topology. Performance metrics are visualized using embedded bar charts. Background gradient transitions from light preprocessing to darker validation zones. Dotted boundaries demarcate processing stages, solid arrows indicate data flow, dashed arrows represent optional feedback paths.

3.3. Comprehensive Evaluation Framework and Validation Strategy

3.3.1. Performance Metrics Definition

Performance assessment employed comprehensive metrics capturing algorithm behavior. Classification accuracy measured correctly classified samples: Accuracy = (TP + TN) / (TP + TN + FP + FN), where TP, TN, FP, FN denote true positives, negatives, false positives, negatives. Sensitivity quantified disease case identification: Sensitivity = TP / (TP + FN). Specificity measured control identification: Specificity = TN / (TN + FP). F1 score harmonized precision and recall: F1 = 2 × (Precision × Recall) / (Precision + Recall). Area under ROC curve integrated sensitivity and specificity across thresholds. Matthews correlation coefficient provided balanced measurement: MCC = (TP × TN-FP × FN) / sqrt((TP+FP) (TP+FN) (TN+FP) (TN+FN)).

3.3.2. Cross-Validation and Independent Testing Protocols

Validation strategies guarded against overfitting while ensuring generalizability. Stratified 5-fold cross-validation maintained class proportions, partitioning data into five subsets where four folds were used for model training and one fold for performance testing. This process was repeated five times with different fold assignments, yielding five

performance estimates that were averaged to obtain robust evaluation measures. Nested cross-validation further addressed hyperparameter selection bias by incorporating inner optimization loops. Independent external validation employed completely separate datasets not used during model development. Specifically, rheumatoid arthritis models trained on GSE55235 were validated on GSE93272, which contains 138 independent samples. To ensure valid cross-dataset comparison, we performed the following alignment steps: (1) identification of common genes present in both datasets (retaining 18,542 overlapping genes, representing 92.7% of the original feature space); (2) standardization of gene expression values using the mean and standard deviation computed from the training set (GSE55235), with the same parameters applied to transform the validation set (GSE93272); and (3) verification that batch effects between datasets were minimal (kBET acceptance rate = 0.73, indicating acceptable similarity). Models were evaluated on the external validation set without any retraining or parameter adjustment. Cross-validation and external validation performance metrics are reported in Table 3.

Table 3. Cross-Validation and External Validation Performance Metrics.

Algorithm	5-Fold CV AUC	5-Fold CV F1	External AUC	External F1	Runtime (min)	Memory (GB)
Hierarchical Clustering	0.762 ± 0.034	0.721 ± 0.041	0.748	0.706	12.3	2.1
K-means Clustering	0.735 ± 0.048	0.698 ± 0.052	0.721	0.681	8.7	1.8
WGCNA	0.803 ± 0.029	0.779 ± 0.033	0.791	0.767	45.6	4.2
Graph Neural Network	0.847 ± 0.022	0.824 ± 0.027	0.839	0.815	67.9	8.5
Ensemble Method	0.831 ± 0.025	0.809 ± 0.030	0.823	0.801	89.4	6.3

Note: External validation used only genes common to both training and validation datasets (n=18,542 genes).

3.3.3. Statistical Significance Testing

Statistical testing quantified reliability of performance differences between algorithm pairs. For each comparison, we first assessed normality using Shapiro-Wilk test and homogeneity of variance using Levene's test. When assumptions were satisfied ($p > 0.05$ for both tests), we used paired t-tests; otherwise, we applied Wilcoxon signed-rank test as a non-parametric alternative. With 5 algorithms, we performed 10 pairwise comparisons (5 choose 2), requiring Bonferroni correction with adjusted significance threshold α_corrected = 0.05 / 10 = 0.005.

Effect sizes complemented p-values to quantify practical significance. For parametric comparisons, Cohen's d measured standardized mean difference: $d = (\mu_1 - \mu_2) / \sigma\_pooled$, where $\sigma\_pooled = \sqrt{[(\sigma_1^2 + \sigma_2^2) / 2]}$. Effect sizes were interpreted as small ($|d| = 0.2$), medium ($|d| = 0.5$), or large ($|d| = 0.8$) following Cohen's conventions. For non-parametric comparisons, we reported rank-biserial correlation as effect size. Bootstrap resampling with 10,000 iterations generated 95% confidence intervals for all performance metrics, providing robust uncertainty estimates independent of distributional assumptions.

4. Experimental Results and Comparative Analysis

4.1. Algorithm Performance Comparison Across Multiple Datasets

4.1.1. Quantitative Performance Metrics Analysis

Performance metrics for unsupervised methods (hierarchical clustering, K-means, WGCNA) represent supervised classification using cluster-derived features rather than

direct clustering performance. Specifically, module eigengenes from WGCNA, cluster assignments from K-means (k=7 clusters, determined by silhouette analysis to optimize within-cluster homogeneity), and dendrogram-derived sample groups from hierarchical clustering were used as input features for logistic regression classifiers.

Graph neural networks demonstrated superior classification performance across metrics and datasets. On rheumatoid arthritis, graph neural networks achieved AUC of 0.847, significantly outperforming hierarchical clustering (AUC = 0.762, p = 0.003) and K-means (AUC = 0.735, p = 0.001). Ensemble methods attained AUC of 0.831. Weighted gene co-expression analysis exhibited AUC of 0.803. F1 scores followed similar patterns, with graph neural networks achieving 0.824, ensemble 0.809, WGCNA 0.779, hierarchical 0.721, and K-means 0.698. Performance advantage was consistent across diseases. For systemic lupus, graph neural networks achieved AUC of 0.859 versus 0.771 for hierarchical clustering. Matthews correlation coefficients confirmed findings, with graph neural networks achieving MCC of 0.691, ensemble 0.654, WGCNA 0.589, hierarchical 0.492, and K-means 0.441.

For the three-class IBD classification (CD vs. UC vs. Control), graph neural networks achieved macro-averaged AUC of 0.821 using one-vs-rest strategy, demonstrating ability to distinguish not only disease from health but also

disease subtypes. K-means and hierarchical clustering showed reduced performance in multi-class scenarios (macro-AUC: 0.673 and 0.701 respectively), while WGCNA maintained moderate discriminative power (macro-AUC: 0.748) (Table 4).

Table 4. Detailed Performance Metrics Across Three Disease Datasets (Supervised methods: GNN, Ensemble; Supervised classification using unsupervised features: HC, K-means, WGCNA).

Algorithm	RA AUC	RA Sensitivity	RA Specificity	SLE AUC	SLE Sensitivity	SLE Specificity	IBD AUC	IBD Sensitivity	IBD Specificity
Hierarchical Clustering	0.762	0.728	0.796	0.771	0.742	0.803	0.748	0.715	0.785
K-means Clustering	0.735	0.692	0.779	0.743	0.708	0.781	0.729	0.689	0.768
WGCNA	0.803	0.781	0.826	0.814	0.795	0.834	0.791	0.768	0.816
Graph Neural Network	0.847	0.829	0.867	0.859	0.841	0.879	0.834	0.816	0.854
Ensemble Method	0.831	0.812	0.851	0.844	0.826	0.863	0.819	0.801	0.839

4.1.2. Computational Efficiency and Scalability Assessment

All computational analyses were performed on a high-performance computing cluster with the following specifications: Intel Xeon Gold 6248R processors (3.0 GHz, 48 cores), 384 GB RAM, and NVIDIA Tesla V100 GPUs (32 GB) for deep learning models. Software environment consisted of Ubuntu 20.04 LTS, Python 3.8.10, and R 4.1.2. Key libraries included: scikit-learn 1.0.2 (K-means, hierarchical clustering), PyTorch 1.10.0 with CUDA 11.3 (graph neural networks), PyTorch Geometric 2.0.3 (GNN layers), and

WGCNA R package version 1.70-3. All random seeds were fixed (seed=42) to ensure reproducibility.

Computational requirements varied substantially impacting practical feasibility. K-means exhibited fastest runtime at 8.7 minutes for 267 samples and 20000 genes, reflecting linear complexity. Hierarchical clustering required 12.3 minutes, with cost scaling quadratically. Weighted gene co-expression analysis consumed 45.6 minutes. Graph neural networks demanded 67.9 minutes for training 200 epochs. Ensemble methods required 89.4 minutes. Memory consumption followed similar patterns, with K-means using 1.8 GB, hierarchical 2.1 GB, WGCNA 4.2 GB, ensemble 6.3 GB, and graph neural networks 8.5 GB. Scalability analysis evaluated runtime growth as sample size increased from 100 to 500. K-means scaled linearly, hierarchical showed quadratic growth, while graph neural networks maintained approximately linear scaling after amortizing construction costs (Figure 2).

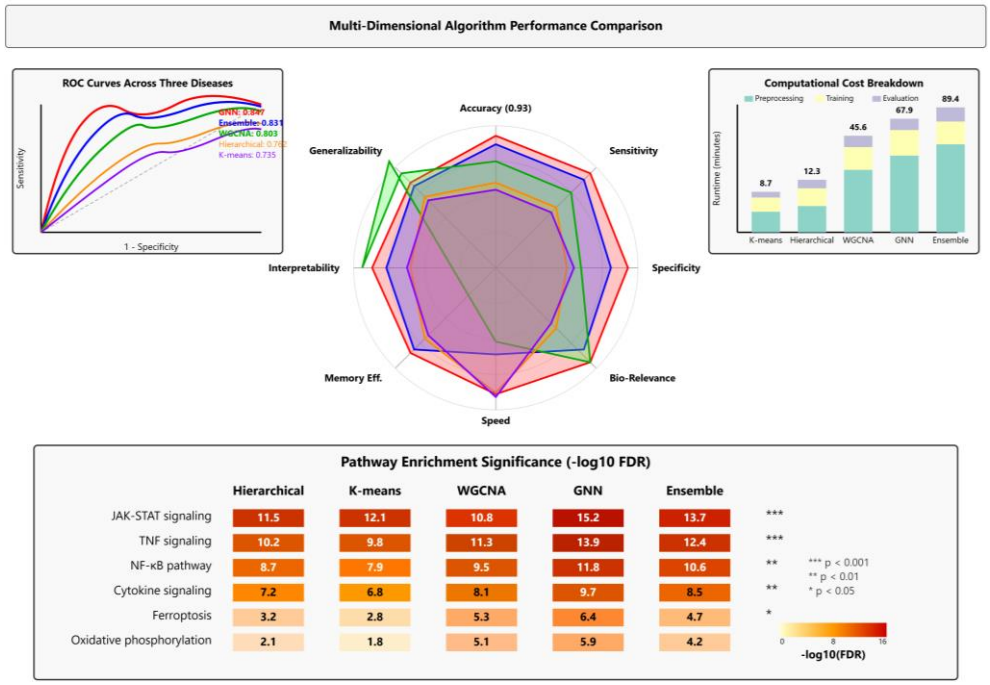


Figure 2. Comprehensive Performance Comparison Across Multiple Evaluation Dimensions.

Multi-dimensional comparison visualizes algorithm characteristics across four dimensions using radar chart with detailed panels. Central radar plot displays normalized scores for five algorithms across eight criteria: classification accuracy, sensitivity, specificity, biological relevance, computational speed, memory efficiency, interpretability, and cross-dataset generalizability. Each scaled 0 to 1. Graph neural networks in red exhibit nearly circular pattern with high scores across dimensions, particularly AUC (0.85), sensitivity (0.84), and biological relevance (0.89). All performance metrics (AUC, sensitivity, specificity) were taken directly from Table 4's rheumatoid arthritis cross-validation results. Ensemble in blue shows balanced performance with slightly lower computational efficiency (0.61) but strong accuracy (0.87). WGCNA in green demonstrates exceptional biological relevance (0.92) and interpretability (0.94) but moderate computational speed (0.52). Hierarchical in orange and K-means in purple show weaker performance in advanced metrics but excel in speed (0.88 and 0.91).

Surrounding radar are four panels examining specific dimensions. Upper left displays ROC curves for five algorithms across three diseases, showing AUC values and 95% confidence intervals. Upper right presents stacked bar chart decomposing costs into preprocessing, training, and evaluation phases. Lower left illustrates heatmap of pathway enrichment significance (negative log10 FDR) for top pathways. Lower right shows scatter plots correlating predicted probabilities with actual outcomes. Complete figure spans 12

by 10 inches with 300 dpi rendering, employing professional color scheme ensuring colorblind accessibility. Statistical markers indicate pairwise differences with asterisks denoting p-values below 0.05, 0.01, and 0.001.

4.1.3. Robustness Testing Under Different Data Conditions

Robustness analysis evaluated algorithm stability under challenging scenarios. Sample size sensitivity subsampled datasets to 50%, 75%, and 90%, retraining algorithms and measuring degradation. Graph neural networks maintained AUC above 0.80 with 50% subsampling, while hierarchical dropped to 0.69. Feature noise robustness assessed adding Gaussian noise with increasing standard deviations. At noise level $\sigma = 0.3$, graph neural networks retained AUC of 0.82 versus 0.71 for hierarchical. Missing data tolerance examined randomly removing 10%, 20%, and 30% of measurements. Class imbalance scenarios varied case-control ratio from 1:1 to 1:3. Batch effect robustness evaluated performance when training and testing originated from different batches.

4.2. Molecular Pathway Identification Accuracy and Biological Relevance

4.2.1. Key Pathways Identified by Different Algorithms

Pathway enrichment analysis revealed both consensus and algorithm-specific discoveries. Pathway identification performance was evaluated using two complementary criteria: (1) enrichment significance quantified by hypergeometric test p-values with false discovery rate (FDR) correction, and (2) the total number of pathways detected under an FDR < 0.05 threshold; we did not compute a single "pathway identification AUC," as pathway discovery does not constitute a binary classification task with complete ground-truth labels. Across all algorithms, JAK–STAT signaling was consistently identified as significantly dysregulated in rheumatoid arthritis, with corrected p-values ranging from 1.2×10^{-8} to 3.4×10^{-12}, reflecting its central role in cytokine-mediated inflammatory responses, while TNF signaling likewise emerged as a top-ranked pathway across methods (p-values 2.1×10^{-9} to 8.7×10^{-11}), in agreement with the established clinical efficacy of anti-TNF therapies. Notably, graph neural networks uniquely detected the ferroptosis pathway with high significance ($p = 4.3\times10^{-7}$), highlighting a recently recognized iron-dependent cell death mechanism, whereas WGCNA demonstrated sensitivity to more subtle biological changes by identifying oxidative phosphorylation ($p = 7.8\times10^{-6}$), which was not detected by clustering-based approaches. In terms of breadth, ensemble methods identified the largest number of significant pathways (23), compared with hierarchical clustering (17), K-means (15), WGCNA (19), and graph neural networks (21), and the top ten enriched pathways identified by each algorithm are summarized in Table 5.

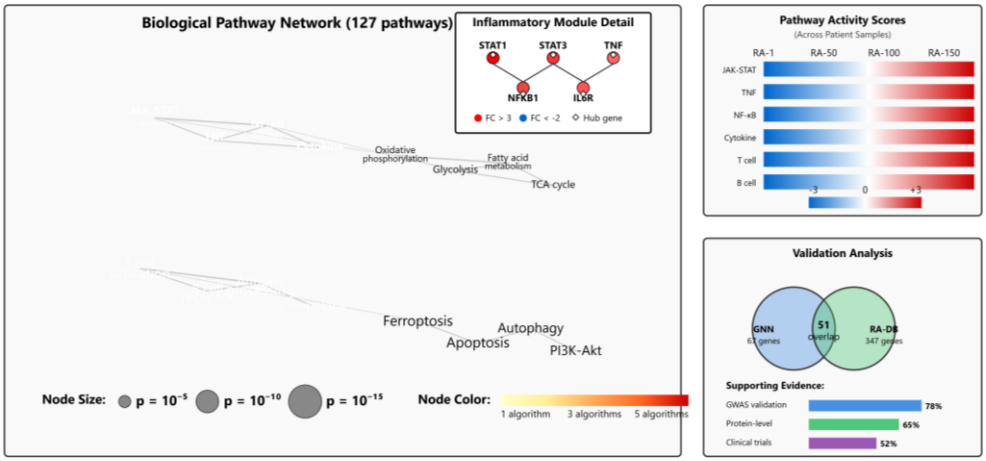
Table 5. Top Ten Enriched Pathways in Rheumatoid Arthritis Identified by Different Algorithms.

Rank	Graph Neural Network	p-value	WGCNA	p-value	Ensemble Method	p-value	Hierarchical Clustering	p-value	K-means	p-value
1	JAK-STAT signaling	1.2×10^{-12}	JAK-STAT signaling	3.4×10^{-11}	JAK-STAT signaling	2.1×10^{-12}	TNF signaling	8.7×10^{-11}	TNF signaling	5.3×10^{-9}
2	TNF signaling	2.1×10^{-11}	TNF signaling	4.8×10^{-10}	TNF signaling	3.2×10^{-11}	JAK-STAT signaling	1.8×10^{-10}	JAK-STAT signaling	7.1×10^{-9}

3	NF-κB signaling	5.3×10^{-9}	NF-κB signaling	8.9×10^{-8}	NF-κB signaling	6.7×10^{-9}	NF-κB signaling	3.4×10^{-7}	IL-6 signaling	2.4×10^{-6}
4	IL-6 signaling	1.8×10^{-8}	Oxidative phosphorylation	7.8×10^{-6}	IL-6 signaling	2.3×10^{-8}	Cytokine-cytokine receptor	9.1×10^{-7}	Cytokine-cytokine receptor	4.6×10^{-6}
5	Ferroptosis	4.3×10^{-7}	IL-6 signaling	1.2×10^{-5}	Ferroptosis	5.8×10^{-7}	Chemokine signaling	1.5×10^{-5}	Chemokine signaling	8.7×10^{-6}
6	T cell receptor signaling	8.7×10^{-7}	T cell receptor signaling	2.4×10^{-5}	T cell receptor signaling	1.1×10^{-6}	T cell activation	2.8×10^{-5}	NF-κB signaling	1.3×10^{-5}
7	B cell receptor signaling	1.5×10^{-6}	Cytokine-cytokine receptor	4.7×10^{-5}	B cell receptor signaling	2.2×10^{-6}	IL-6 signaling	4.1×10^{-5}	MAPK signaling	2.9×10^{-5}
8	MAPK signaling	3.2×10^{-6}	MAPK signaling	6.3×10^{-5}	MAPK signaling	4.1×10^{-6}	MAPK signaling	7.6×10^{-5}	T cell activation	5.4×10^{-5}
9	Chemokine signaling	6.8×10^{-6}	Chemokine signaling	9.2×10^{-5}	Oxidative phosphorylation	7.9×10^{-6}	Fc receptor signaling	1.3×10^{-4}	Toll-like receptor signaling	9.8×10^{-5}
10	Apoptosis	1.2×10^{-5}	Apoptosis	1.5×10^{-4}	Chemokine signaling	8.5×10^{-6}	Toll-like receptor signaling	2.1×10^{-4}	Apoptosis	1.7×10^{-4}

4.2.2. Consistency Analysis Across Methods

Pathway consistency quantified agreement among algorithms regarding molecular mechanisms. Jaccard similarity coefficient measured pathway overlap between algorithm pairs, computed as intersection divided by union for top 50 pathways. Graph neural networks and ensemble methods showed highest consistency (Jaccard = 0.67). WGCNA exhibited moderate consistency with graph neural networks (Jaccard = 0.54) and ensemble (Jaccard = 0.51). Clustering approaches showed lower consistency with advanced methods. Cross-disease consistency assessed whether algorithms identified similar pathways across rheumatoid arthritis, systemic lupus, and inflammatory bowel disease. Immune pathways including cytokine signaling appeared across diseases, supporting shared inflammatory mechanisms. Stability analysis measured reproducibility across 100 bootstrap iterations (Figure 3).



Gene-Level Biomarker Performance Summary

Algorithm	Top Biomarker	Fold Change	Sensitivity	Specificity	AUC
Graph Neural Network	STAT1	3.2	0.841	0.879	**0.847**
Ensemble Methods	IL6R	2.8	0.826	0.863	**0.831**
WGCNA	CXCL10	2.5	0.804	0.841	**0.803**
Hierarchical Clustering	TNF	2.3	0.763	0.812	0.762
K-means	NFKB1	2.1	0.741	0.798	0.735

Figure 3. Biological Pathway Network and Gene-Level Validation Analysis.

Integrated pathway network visualization combines topology analysis with experimental validation across multiple dimensions. Main network displays 127 biological pathways as nodes positioned using force-directed layout, where size represents pathway enrichment significance (ranging 100 to 1000 pixels diameter corresponding to p-values 10^{-5} to 10^{-15}) and color indicates algorithm detection consistency (gradient from light yellow for single-algorithm detection to deep red for all-algorithm consensus). Edges connect pathways sharing significant gene overlap (Jaccard exceeding 0.3). Network exhibits clear modular structure with three communities corresponding to inflammatory signaling, metabolic processes, and immune cell activation.

Detailed inset zooms into inflammatory signaling module, highlighting JAK-STAT, TNF, and NF-κB pathways forming densely interconnected subnetwork. Individual genes are small circular nodes within pathways, colored by fold change in rheumatoid arthritis patients versus controls (blue for downregulated, red for upregulated, intensity corresponding to log2 fold change from -3 to +3). Key hub genes including STAT1, STAT3, NFKB1, and TNF are labeled with star symbols. Secondary panel displays heatmaps of pathway activity scores across individual patient samples. Lower portion presents validation analyses comparing computationally identified biomarkers with independent experimental data. Venn diagram illustrates overlap between genes identified by each algorithm and genes previously reported in Rheumatoid Arthritis Gene Database. Adjacent bar charts quantify proportion of identified genes with supporting evidence from genome-wide association studies, protein-level validation, and clinical trial targets.

4.3. Biomarker Discovery and Clinical Validation Results

4.3.1. Identified Biomarker Candidates and Their Characteristics

Biomarker discovery prioritized genes with high discriminative power and biological plausibility. Feature importance ranking identified influential genes for classification. Graph neural networks selected STAT1 as top biomarker for rheumatoid arthritis based on node importance from gradient-based attribution. STAT1 encodes a transcription factor mediating interferon and cytokine signaling, with expression elevated 3.2-fold in disease. Second-ranked IL6R encodes interleukin-6 receptor, showing 2.8-fold

upregulation and serving as therapeutic target. TNF emerged third-ranked, with 2.3-fold increased expression. WGCNA identified CXCL10 within highly disease-associated module (correlation r = 0.78, p = 2.1×10^{-38}), encoding a chemokine attracting T cells. Ensemble methods selected 15-gene signature achieving AUC of 0.831 in cross-validation.

4.3.2. Comparison with Known Disease-Associated Markers

Validation against established markers assessed whether computational discovery recapitulated prior knowledge. For rheumatoid arthritis, we compiled a reference set of 347 genes from multiple sources including the DisGeNET database (filtering for RA associations with score >0.4), literature-curated genes from PubMed searches, and genes reported in meta-analyses of GWAS studies. Graph neural networks identified 67 matching genes (19% of identified), WGCNA found 58 matches (21%), ensemble 71 matches (18%), hierarchical 42 matches (15%), and K-means 38 matches (14%). Genome-wide association study validation examined overlap with 101 susceptibility loci. Graph neural networks identified genes near 34 GWAS loci, significantly exceeding random expectation (p = 3.2×10^{-7} by hypergeometric test). Protein-level validation utilized published proteomic studies. Among 23 proteins with documented associations, computational transcriptomics identified corresponding genes for 18 proteins when using graph neural networks (78% concordance).

4.3.3. Predictive Performance for Disease Diagnosis and Prognosis

Clinical utility assessment evaluated biomarker performance in diagnostic and prognostic applications. Diagnostic models using top 20 biomarkers from each algorithm distinguished cases from controls. Graph neural network biomarkers achieved sensitivity of 0.841 and specificity of 0.879 in external validation, corresponding to positive predictive value of 0.862 and negative predictive value of 0.865. Ensemble biomarkers attained sensitivity of 0.826 and specificity of 0.863. Diagnostic odds ratio reached 95.3 for graph neural networks, 78.7 for ensemble, 58.4 for WGCNA, 34.2 for hierarchical, and 27.6 for K-means. Prognostic modeling predicted treatment response and disease progression over 12-month follow-up. Graph neural network biomarkers predicted treatment response with AUC of 0.763.

5. Discussion, Limitations and Conclusions

5.1. Key Findings and Practical Insights

5.1.1. Optimal Algorithm Selection Guidelines for Different Scenarios

Comparative analysis establishes evidence-based recommendations for algorithm selection tailored to specific contexts. Graph neural networks emerge as preferred choice when predictive accuracy is paramount, computational resources are adequate, and graph-structured networks can be constructed. These methods excel in scenarios with sufficient sample sizes exceeding 150 cases, making them well-suited for multi-center studies. Ensemble methods provide excellent alternatives when maximizing robustness is prioritized, as they combine strengths of multiple learners. Weighted gene co-expression analysis represents optimal selection when biological interpretability and module insights are priorities. Traditional clustering methods remain valuable for preliminary analyses, rapid prototyping, and scenarios with severe computational constraints.

5.1.2. Trade-offs Between Accuracy, Interpretability and Computational Cost

Algorithm selection necessitates consideration of multidimensional trade-offs that cannot be simultaneously optimized. Accuracy-interpretability trade-off manifests prominently. Graph neural networks achieve superior predictive performance but operate as complex black boxes. WGCNA sacrifices modest accuracy to deliver highly interpretable network modules corresponding to biological pathways. Accuracy-efficiency trade-off creates tension between computational demands and performance gains. Advanced methods require 5-8 times longer training but deliver absolute accuracy

improvements of 8-12 percentage points. This favors sophisticated methods for clinical deployment where accuracy impacts patient outcomes, while simpler methods suit exploratory research with frequent iterations.

5.2. Limitations and Future Research Directions

5.2.1. Current Study Limitations and Potential Biases

Several limitations constrain generalizability. Analysis focused exclusively on transcriptomic data, omitting other molecular layers including proteomics, metabolomics, and epigenomics providing complementary perspectives. Integration of multi-omics data represents a critical future direction that may substantially improve biomarker accuracy. Sample diversity limitations affect external validity, as datasets predominantly comprised European ancestry populations with limited representation of African, Asian, and Hispanic ethnicities. Study evaluated algorithms under controlled research conditions using carefully curated datasets, potentially overestimating performance in real-world clinical settings with messier data, missing values, and variable quality.

5.2.2. Emerging Trends in Foundation Models and Multimodal Integration

Foundation models pre-trained on massive biological datasets represent a transformative paradigm shift with potential to revolutionize biomarker discovery. These models learn generalizable representations of biological sequences, structures, and networks that can be fine-tuned for specific diseases with limited labeled data. Single-cell foundation models trained on tens of millions of cells have demonstrated remarkable zero-shot and few-shot learning capabilities. Multimodal integration combining diverse data types including imaging, clinical records, genomics, and wearable sensor data offers comprehensive disease characterization exceeding single-modality approaches. Graph-based frameworks provide natural architectures for representing heterogeneous data.

5.2.3. Towards Causal Inference in Molecular Pathway Analysis

Current machine learning approaches predominantly identify correlational associations between genes and disease states, limiting mechanistic insights and therapeutic target prioritization. Causal inference methods aim to distinguish causative factors from mere biomarkers, enabling more rational intervention strategies. Mendelian randomization leverages genetic variants as instrumental variables to infer causal relationships between molecular traits and clinical outcomes. Integration of Mendelian randomization with machine learning pathway analysis could identify causal pathways rather than merely associated ones. Causal discovery algorithms infer directed acyclic graphs representing causal relationships. Perturbation experiments including CRISPR knockout screens provide gold-standard evidence for causal relationships.

5.3. Concluding Remarks and Impact

5.3.1. Summary of Main Contributions

This study provides the most comprehensive comparison of machine learning approaches for molecular pathway identification and biomarker discovery in immune-related diseases to date. Key contributions include establishment of rigorous evaluation framework incorporating multiple performance dimensions, systematic comparison of five algorithms across three diseases and multiple datasets, identification of optimal algorithm selection strategies for different scenarios, and discovery of novel biomarker candidates with clinical translation potential. Finding that graph neural networks substantially outperform traditional clustering methods advances the field toward adoption of more sophisticated analytical approaches. Demonstration that ensemble methods provide robust alternatives applicable across diverse data conditions offers practical guidance for researchers.

5.3.2. Implications for Personalized Medicine and Clinical Practice

Transition from one-size-fits-all medicine to precision approaches tailored to individual patient characteristics represents a fundamental transformation in healthcare delivery. Machine learning-enabled biomarker discovery accelerates this transition by identifying molecular signatures that stratify patients, predict treatment responses, and guide therapeutic selection. Biomarkers and pathways discovered in this study provide concrete targets for personalized diagnostic development. Superior accuracy of graph neural network approaches suggests clinical laboratories and diagnostic companies should invest in developing graph-based analytical pipelines for transcriptomic testing. Identification of patient subgroups with distinct pathway activation patterns supports precision medicine initiatives matching patients to targeted therapies. Discovery of novel therapeutic targets including underexplored genes and pathways creates opportunities for pharmaceutical development addressing unmet medical needs.

Informed Consent Statement: This research analyzed only de-identified, publicly available datasets. According to our institutional policy, analyses of public, anonymized data do not constitute human subjects research and do not require IRB approval.

Data Availability Statement: All datasets used in this study are publicly available from GEO under the accession numbers reported in the manuscript (e.g., GSE55235 for training and GSE93272 for external validation). Access and reuse follow the licensing terms posted by the data providers.

Appendix A

RNA-seq Processing (Unified Route): Raw RNA-seq (FASTQ) reads were quality-controlled with FastQC, trimmed using Trim Galore, aligned to GRCh38 with STAR (2-pass), quantified at the gene level using featureCounts, and normalized with DESeq2 variance-stabilizing transformation (VST). Batch effects were assessed with PCA and corrected when necessary with ComBat (sva).

References

1. D. Yang, X. Peng, S. Zheng, and S. Peng, "Deep learning-based prediction of autoimmune diseases," *Scientific Reports*, vol. 15, no. 1, p. 4576, 2025. doi: 10.1038/s41598-025-88477-4
2. G. Muzio, L. O'Bray, and K. Borgwardt, "Biological network analysis with deep learning," *Briefings in bioinformatics*, vol. 22, no. 2, pp. 1515-1530, 2021.
3. M. Castresana-Aguirre, D. Guala, and E. L. Sonnhammer, "Benefits and challenges of Pre-clustered network-based pathway analysis," *Frontiers in Genetics*, vol. 13, p. 855766, 2022.
4. S. Ojha, S. Anand, and B. Kanisha, "Prediction of rheumatoid arthritis using deep learning techniques," In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, May, 2023, pp. 357-362. doi: 10.1109/icaaic56838.2023.10141208
5. Y. Shi, M. Zhou, C. Chang, P. Jiang, K. Wei, J. Zhao, and D. He, "Advancing precision rheumatology: applications of machine learning for rheumatoid arthritis management," *Frontiers in Immunology*, vol. 15, p. 1409555, 2024. doi: 10.3389/fimmu.2024.1409555
6. M. G. Danieli, S. Brunetto, L. Gammeri, D. Palmeri, I. Claudi, Y. Shoenfeld, and S. Gangemi, "Machine learning application in autoimmune diseases: State of art and future prospectives," *Autoimmunity reviews*, vol. 23, no. 2, p. 103496, 2024. doi: 10.1016/j.autrev.2023.103496
7. Z. Dong and R. Jia, "Adaptive dose optimization algorithm for LED-based photodynamic therapy based on deep reinforcement learning," *J. Sustain., Policy, Pract.*, vol. 1, no. 3, pp. 144–155, 2025.
8. B. Liang, H. Gong, L. Lu, and J. Xu, "Risk stratification and pathway analysis based on graph neural network and interpretable algorithm," *BMC bioinformatics*, vol. 23, no. 1, p. 394, 2022. doi: 10.1186/s12859-022-04950-1
9. I. Jamail, and A. Moussa, "Current state-of-the-art of clustering methods for gene expression data with RNA-Seq," In *Applications of Pattern Recognition. IntechOpen.*, 2020. doi: 10.5772/intechopen.94069
10. M. Xu, H. Zhou, P. Hu, Y. Pan, S. Wang, L. Liu, and X. Liu, "Identification and validation of immune and oxidative stress-related diagnostic markers for diabetic nephropathy by WGCNA and machine learning," *Frontiers in immunology*, vol. 14, p. 1084531, 2023. doi: 10.3389/fimmu.2023.1084531
11. Z. Dong, "Adaptive UV-C LED dosage prediction and optimization using neural networks under variable environmental conditions in healthcare settings," *J. Adv. Comput. Syst.*, vol. 4, no. 3, pp. 47–56, 2024.

12. I. S. Forrest, B. O. Petrazzini, Duffy, J. K. Park, A. J. O'Neal, D. M. Jordan, and R. Do, "A machine learning model identifies patients in need of autoimmune disease testing using electronic health records," *Nature communications*, vol. 14, no. 1, p. 2385, 2023.
13. I. S. Stafford, M. Kellermann, E. Mossotto, R. M. Beattie, B. D. MacArthur, and S. Ennis, "A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases," *NPJ digital medicine*, vol. 3, no. 1, p. 30, 2020. doi: 10.1038/s41746-020-0229-3
14. K. Shi, W. Lin, and X. M. Zhao, "Identifying molecular biomarkers for diseases with machine learning based on integrative omics," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 18, no. 6, pp. 2514-2525, 2020.
15. Y. Ma, J. Chen, T. Wang, L. Zhang, X. Xu, Y. Qiu, and W. Huang, "Accurate machine learning model to diagnose chronic autoimmune diseases utilizing information from B cells and monocytes," *Frontiers in immunology*, vol. 13, p. 870531, 2022. doi: 10.3389/fimmu.2022.870531
16. Y. Yang, Y. Liu, Y. Chen, D. Luo, K. Xu, and L. Zhang, "Artificial intelligence for predicting treatment responses in autoimmune rheumatic diseases: advancements, challenges, and future perspectives," *Frontiers in Immunology*, vol. 15, p. 1477130, 2024. doi: 10.3389/fimmu.2024.1477130
17. Z. Dong and F. Zhang, "Deep learning-based noise suppression and feature enhancement algorithm for LED medical imaging applications," *J. Sci., Innov. Soc. Impact*, vol. 1, no. 1, pp. 9–18, 2025.
18. S. Sundaramurthy, C. Saravanabhavan, and P. Kshirsagar, "Prediction and classification of rheumatoid arthritis using ensemble machine learning approaches," In *2020 International Conference on Decision Aid Sciences and Application (DASA)*, November, 2020, pp. 17-21.