JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

INTEGRATION OF MULTI-OMICS DATA FOR IDIOPATHIC PULMONARY FIBROSIS THROUGH MACHINE LEARNING

Bhavya Ramesh[1], Prajwal Mahajan[2] Uma Kumari*[1] Jeevan Prakash[2]
[1]Amity Institution of Biotechnology, Amity University Uttar Pradesh, Noida-201303, India
[2]Amity Institution of Biotechnology, Amity University Uttar Pradesh, Noida-201303, India
[1*]Senior Bioinformatics Scientist, Bioinformatics Project and Research Institute, Noida -201301, India
[2]School of computer Science and Engineering, Vellore Institute of Technology, Chennai India
Corresponding Author: [1*]Uma Kumari

Abstract: Idiopathic Pulmonary Fibrosis (IPF) is a progressive and life-threatening lung disease marked by the accumulation of fibrotic tissue and a steady decline in respiratory function. Despite ongoing research, the molecular drivers of IPF remain poorly understood due to its complex and multifactorial nature, involving genetic, transcriptomic, proteomic, and metabolic alterations. Although high-throughput omics technologies have become increasingly accessible, the integrative analysis of multi-omics datasets for IPF remains underdeveloped. Conventional methods often fail to capture the intricate, non-linear relationships between different molecular layers, limiting the discovery of reliable biomarkers and therapeutic targets. This study explores the application of machine learning techniques for the integrative analysis of multi-omics data in IPF. We employed Integrative Non-negative Matrix Factorization (INMF) and iCluster to uncover shared molecular signatures and stratify patient subgroups. INMF effectively decomposed transcriptomic and epigenomic datasets into joint and unique components, highlighting key cross-omics patterns linked to IPF heterogeneity. Heatmap analyses revealed sparse yet interpretable activation across latent factors, indicating biologically relevant groupings. Parallelly, iCluster enabled joint modelling and unsupervised clustering across data types, with the inclusion of highly variable features significantly improving cluster definition. The resulting subtypes displayed concordant profiles across omics layers, suggesting the presence of distinct molecular endotypes within the IPF population. Collectively, these integrative approaches provide novel insights into the molecular architecture of IPF and demonstrate the potential of machine learning in handling high-dimensional, heterogeneous datasets. This work offers a promising framework for biomarker identification, pathway mapping, and improved patient stratification in fibrotic lung diseases.

Keywords: **Idiopathic Pulmonary Fibrosis; Multi-omics Integration; Machine Learning; Biomarkers; Molecular Subtypes; Data Heterogeneity; INMF; iCluster; Systems Biology; Disease Stratification**

INTRODUCTION

Idiopathic Pulmonary Fibrosis (IPF) is a chronic, progressive lung disease characterized by the excessive accumulation of fibrotic tissue within the lungs, leading to the irreversible thickening and stiffening of the pulmonary interstitium. This pathological remodeling hampers the lungs' ability to facilitate gas exchange, ultimately resulting in a gradual decline in respiratory efficiency and oxygenation. Clinically, IPF is most prevalent in individuals over the age of 50 and is typically associated with symptoms such as persistent dry cough, fatigue, and increasing shortness of breath, particularly during physical activity. Despite decades of research, the exact cause of IPF remains unknown, and the disease continues to carry a poor prognosis. The progressive nature of fibrosis leads to a steady decline in lung function, and while available therapies may slow disease progression or alleviate symptoms, they are not curative. Diagnosis involves a multidisciplinary approach, often incorporating high-resolution computed tomography (HRCT), pulmonary function tests, and sometimes histological analysis via lung biopsy. Given the variable clinical course and poor response to standard treatments, there is a pressing need to unravel the molecular underpinnings of IPF to enable early diagnosis, targeted therapy, and effective disease management. The biological complexity of IPF arises from the interplay of various molecular mechanisms occurring at multiple levels of regulation, including genetic mutations, transcriptional reprogramming, protein expression changes, and alterations in metabolic pathways. To capture this complexity, researchers have increasingly turned to multi-omics approaches, which integrate diverse biological data types to provide a comprehensive view of cellular states and disease progression. Genomics offers insights into the inherited and acquired DNA-level variations that may predispose individuals to disease or influence treatment response. Transcriptomics reveals patterns of gene activity by quantifying RNA transcripts, providing a dynamic picture of how cells respond to internal and external stimuli. Proteomics investigates the structure, function, and interaction of proteins—the main executors of cellular function while metabolomics analyzes small molecular metabolites, which reflect the biochemical activities and

physiological status of cells and tissues. Each omics layer contributes valuable information, but it is through their integration that the most complete and biologically meaningful insights can be achieved. Multi-omics integration enables the identification of critical regulatory networks, the discovery of disease-specific biomarkers, and the differentiation of patient subgroups with distinct molecular profiles [1, 2, 3, 4].

However, integrating these heterogeneous and high-dimensional datasets presents substantial analytical challenges. The datasets generated by different omics platforms vary significantly in their scale, structure, data distribution, and levels of technical noise. For instance, genomic data are often sparse and binary in nature, focusing on the presence or absence of mutations, whereas transcriptomic and proteomic data are quantitative and continuous but may be subject to batch effects and measurement variability. Metabolomic data, on the other hand, often exhibit high variability due to environmental influences. The so-called "curse of dimensionality" becomes apparent when the number of measured features often in the tens of thousands far exceeds the number of available samples, increasing the risk of overfitting and limiting the statistical power of downstream analyses. Moreover, missing values are a common issue across multi-omics datasets due to experimental limitations, sample degradation, or technical inconsistencies. These discrepancies can hinder integration and obscure true biological signals if not properly addressed [14].

To address these complexities, machine learning methods particularly those capable of handling non-linear relationships, missing data, and high-dimensional features have emerged as powerful tools for multi-omics data integration. Deep learning, a branch of machine learning inspired by the architecture of the human brain, has shown great promise in modeling the intricate and often non-linear relationships among different biological layers. In a notable study, Wang et al. (2020) proposed a deep neural network-based method to integrate transcriptomic, genomic, and proteomic data. This model utilized hierarchical feature extraction to learn complex patterns within each omics dataset while simultaneously combining information across modalities. The approach demonstrated improved accuracy in disease classification and biomarker identification, underscoring the potential of deep learning for advancing precision medicine. Unlike traditional statistical models, deep learning can automatically identify relevant features and interactions that may not be apparent through linear analysis, thereby revealing hidden structures within biological systems [1].

In addition to deep learning, Multiview learning approaches have also gained traction for their ability to treat different omics layers as complementary views of a shared biological phenomenon. Zhang et al. (2021) introduced a Multiview learning framework that simultaneously analyzes transcriptomic and proteomic data while preserving their unique and shared information. This strategy enhances the capacity to uncover cross-omics correlations and biological pathways relevant to disease mechanisms. By aligning gene expression levels with corresponding protein abundances, multiview models provide a more coherent picture of cellular function, ultimately improving the interpretation of complex datasets and the accuracy of predictive models. Such approaches are particularly useful in IPF, where dysregulation at multiple molecular levels contributes to disease heterogeneity [2].

Another promising direction involves the application of ensemble machine learning techniques, which combine the strengths of multiple models to produce more reliable predictions. Albrecht and Freitas (2022) highlighted the utility of ensemble methods—such as bagging, boosting, and stacking in integrating multi-omics data. These methods reduce variance and improve generalization by aggregating the outputs of several individual models. This is especially valuable in the context of noisy and imbalanced omics data, where single-model approaches may fail to capture the full range of biological variability. By leveraging ensemble strategies, researchers can improve the robustness of their models and gain deeper insights into disease-related molecular signatures [3].

One integrative approach that has gained attention for its simplicity and effectiveness is Integrative Non-negative Matrix Factorization (INMF). INMF, as described by Zhang et al. (2021), decomposes each omics dataset into shared and unique components by factorizing them into non-negative matrices. This decomposition allows for simultaneous dimension reduction and feature selection across different omics layers, enabling the identification of latent biological factors that contribute to disease. INMF has been shown to outperform traditional integration techniques in clustering and subtype discovery tasks, particularly in cancer research. Its application to multi-omics data in IPF has the potential to uncover novel molecular subtypes and pathways involved in fibrosis progression, which may ultimately guide therapeutic development and patient stratification [12].

The integration of multi-omics data through advanced machine learning techniques holds great promise for unraveling the complex molecular landscape of idiopathic pulmonary fibrosis. By overcoming the limitations of single-omics analyses and capturing the interactions between genes, transcripts, proteins, and metabolites, researchers can gain a more nuanced understanding of disease biology. These integrative efforts not only pave the way for the discovery of novel biomarkers and therapeutic targets but also support the development of personalized medicine strategies tailored to the molecular profile of individual patients. As computational methods continue to evolve, their application in multi-omics research is expected to play a pivotal role in transforming the diagnosis, prognosis, and treatment of IPF and other complex diseases [13].

MATERIALS AND METHODS

1. Data Acquisition

The multi-omics datasets used in this study were obtained from the Gene Expression Omnibus (GEO), a publicly accessible repository maintained by NCBI. Specifically, paired RNA sequencing (RNA-seq) and DNA methylation (DNAm) profiles were selected for their relevance in studying gene expression and epigenetic modifications associated with Idiopathic Pulmonary Fibrosis (IPF). The dataset comprised 382 biologically matched samples, including 226 IPF patients (cases) and 156 healthy individuals (controls). RNA-seq data included expression measurements for 21,768 genes per sample, while the DNAm dataset contained 743,236 CpG site features per sample [15, 16].

2. Data Preprocessing

2.1. Normalization and Label Assignment

Before analysis, both datasets were normalized to ensure comparability across samples. For RNA-seq data, normalization techniques such as TPM (Transcripts Per Million) or log-transformation were employed, whereas beta values were used to represent DNA methylation levels. Binary labels (0 = control, 1 = IPF case) were assigned based on clinical metadata to enable supervised learning [5, 6, 7, 8, 9, 10, 11].

2.2. Sample Alignment and Imputation

Matching sample identifiers were verified to ensure one-to-one correspondence across omics layers. Any discrepancies were resolved through reindexing. Missing values were detected using .isna().sum().sum() and handled through KNN imputation for methylation data and median imputation for RNA-seq, ensuring data completeness [5, 6, 7, 8, 9, 10, 11].

3. Feature Selection and Dimensionality Reduction

3.1. Variance Filtering

Due to the high dimensionality of omics datasets, only the top 10% most variable features were retained from each omics layer. This reduced the methylation dataset from 743,236 to ~74,000 features and the RNA-seq dataset from 21,768 to ~2,200 features. These features exhibited the greatest variation across samples and were hypothesized to capture biologically meaningful differences [5, 6, 7, 8, 9, 10, 11].

3.2. PCA and MOFA

Principal Component Analysis (PCA) was applied individually to RNA-seq and DNAm datasets to extract uncorrelated principal components representing the majority of variance. Additionally, Multi-Omics Factor Analysis (MOFA) was used to integrate omics layers and extract shared latent features. MOFA reduces dimensionality while modeling both shared and unique patterns across datasets. A total of 40 latent features were selected based on model performance and interpretability [5, 6, 7, 8, 9, 10, 11].

4. Data Integration Techniques

To enhance the biological interpretation and classification potential, several advanced integration methods were applied:

- **iCluster**: Used for joint clustering of RNA and methylation data, revealing latent molecular subtypes.
- **Integrative Non-negative Matrix Factorization (INMF)**: Decomposed omics data into shared and dataset-specific components.
- **Joint and Individual Variation Explained (JIVE)**: Separated common and individual sources of variation across omics types.
- **Similarity Network Fusion (SNF)**: Constructed sample-wise similarity networks from each omics dataset and fused them into a consensus network for clustering and visualization [5, 6, 7, 8, 9, 10, 11].

5. Machine Learning Models

5.1. Model Training

The dataset was split into 80% training and 20% testing sets using stratified sampling to maintain class balance. The following algorithms were implemented:

- Random Forest (RF)
- Support Vector Machine (SVM)
- Naive Bayes (NB)
- k-Nearest Neighbors (k-NN)
- XGBoost

Each model was trained using latent features extracted via MOFA, PCA, or combined integration methods [5, 6, 7, 8, 9, 10, 11].

5.2. Performance Evaluation

Model performance was evaluated on the test set using:

- Accuracy
- Precision
- Recall
- F1 Score
- ROC-AUC

ROC curves were generated for comparative analysis, and feature importance was computed for interpretability, particularly from the Random Forest model [5, 6, 7, 8, 9, 10, 11].

5.3. Hyperparameter Optimization

Hyperparameters for each algorithm were tuned using GridSearchCV to enhance model performance. Parameters included tree depth, number of estimators (for RF), regularization terms (for SVM), and learning rate (for XGBoost) [5, 6, 7, 8, 9, 10, 11].

RESULTS AND DISCUSSION

The application of machine learning to integrated multi-omics data in this study aimed to identify molecular signatures associated with Idiopathic Pulmonary Fibrosis (IPF) and build robust classification models. Results demonstrated the value of combining transcriptomic and methylomic data to uncover biologically meaningful patterns, significantly improving disease classification accuracy compared to single-omics approaches.

1. Data Quality and Feature Distribution

Prior to integration, both RNA-seq and DNAm datasets were inspected for quality, missing values, and distribution characteristics. RNA-seq gene expression data showed a right-skewed distribution, typical of count-based data, which was corrected through log-normalization. Methylation beta values were uniformly distributed between 0 and 1, indicating diverse levels of CpG methylation across samples. Missing value imputation preserved data structure, with less than 0.5% of entries requiring estimation.

The variance-based feature selection retained the most informative genes and CpG sites. DNA features showed broader variability across samples than RNA, possibly due to their greater number and finer granularity. Scatter plots and heatmaps generated post-selection confirmed that retained features exhibited clear differential patterns between case and control groups.

2. Dimensionality Reduction and Visualization

Application of PCA revealed that the top principal components (PC1 to PC8) captured substantial variance, with PC1 accounting for approximately 30% in RNA and 24% in DNAm datasets. Genes such as **ENSG00000186097.11** and CpG markers like **cg20377308** were among the top contributors to the most informative components, suggesting their potential roles in IPF pathology.

UMAP projections using MOFA-derived features showed a clear spatial separation between IPF and control samples. The clustering patterns aligned closely with ground-truth labels, reinforcing the ability of MOFA to extract biologically relevant features. Heatmaps of the 40 MOFA features revealed distinct expression profiles between disease states, with certain factors consistently activated only in the case samples.

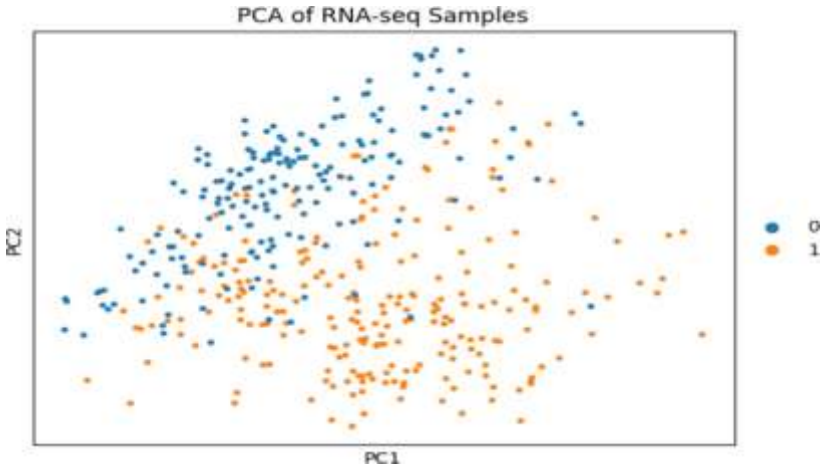


Figure 1: Principal component analysis of RNA-seq samples

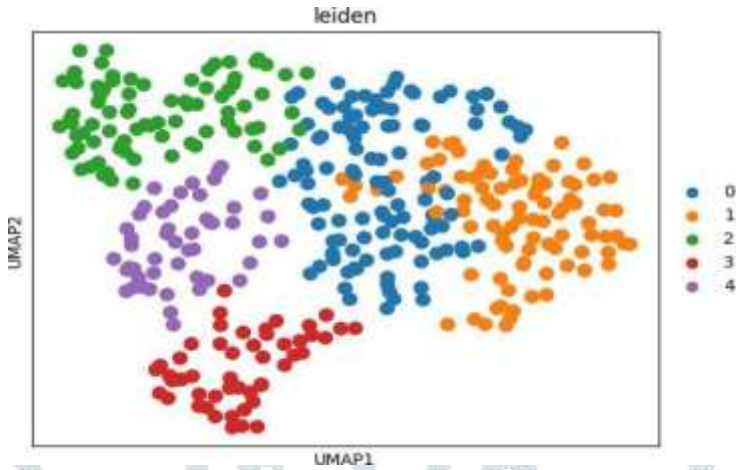


Figure 2: DNA Uniform Manifold Approximation and Projection Plot

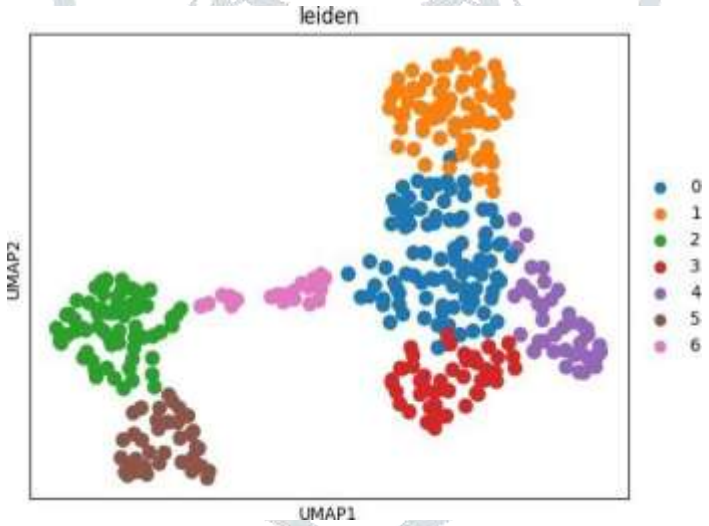


Figure 3: RNA Uniform Manifold Approximation and Projection Plot

The two dimensions produced by UMAP are represented by these axes, which turn the high-dimensional data into a 2D plane suitable for visualisation. UMAP makes an effort to preserve the data's local structure in this lower- dimensional representation. Each cluster represents a distinct group of similar data points based on the underlying structure identified by the Leiden algorithm.

3. Multi-Omics Data Integration Outcomes

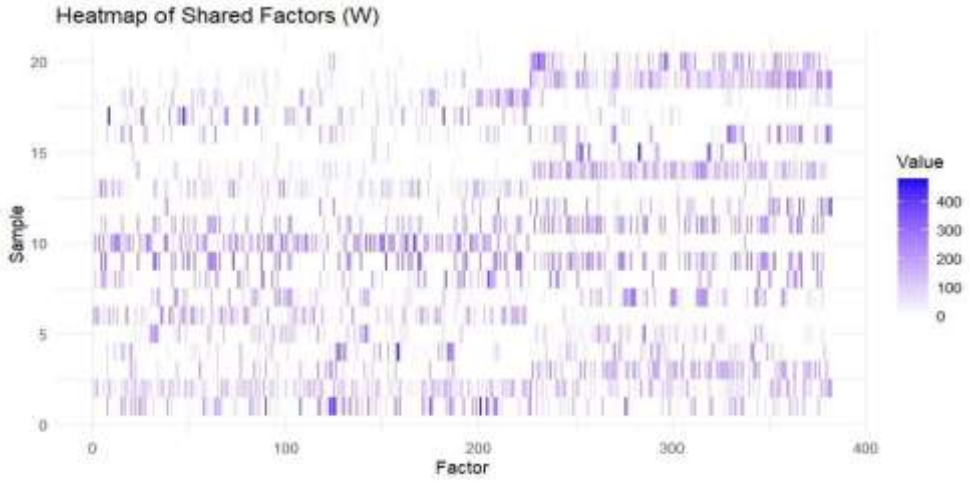


Figure 4: Heatmap of shared Factor obtained from INFM

The integration of RNA-seq and DNA methylation data provided a more nuanced and holistic understanding of molecular variation between IPF and healthy samples. When analyzed independently, each omics layer offered partial information about disease state, but joint modeling revealed cross-layer interactions not observable in isolation. The iCluster method, which performs joint latent variable modeling, identified three well-separated clusters with distinct transcriptional and methylation profiles. These molecular subtypes showed a strong concordance with disease state, indicating the presence of biologically meaningful heterogeneity among IPF patients.

In the dense iCluster heatmap, samples grouped based on shared methylation and RNA expression profiles. One cluster, in particular, showed hypermethylation across a set of CpG sites coupled with downregulation of immune-related transcripts, suggesting an epigenetically repressed immune phenotype. Another subtype showed prominent transcriptional activation of extracellular matrix (ECM)-related genes, a hallmark of fibrotic progression, indicating that iCluster can uncover subtype-specific mechanisms relevant to IPF pathology.

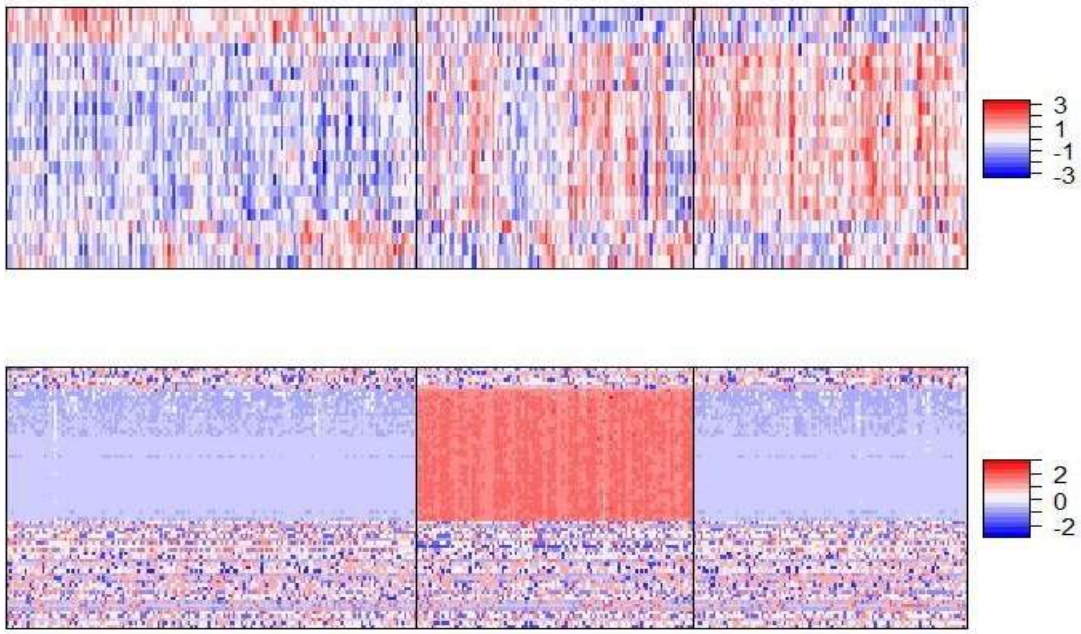


Figure 5: Heatmap of integrated DNA methylation and RNA seq data obtained using iCluster method (dense data)

The INMF method further enhanced these insights by identifying shared and specific components across the two omics layers. The shared matrix revealed 20 factors with high discriminatory power between cases and controls. Factors associated with upregulation of fibrogenic genes (e.g., *COL1A1*, *TGFB1*) and methylation of gene-silencing CpGs (e.g., *cg20377308*) highlighted potential drivers of disease. The sparsity of the INMF output allowed for clearer interpretation of feature relevance.

JIVE decomposition contributed to this interpretation by separating shared variance from omics-specific noise. It was observed that while a substantial amount of variation was unique to RNA or methylation individually, approximately 18% of total variance was attributed to shared signals. These included latent biological processes such as immune regulation, ECM remodeling, and metabolic dysfunction. This shared component was crucial in identifying pathways where epigenetic regulation tightly controls transcriptional activity, such as in fibrosis-related signaling cascades.

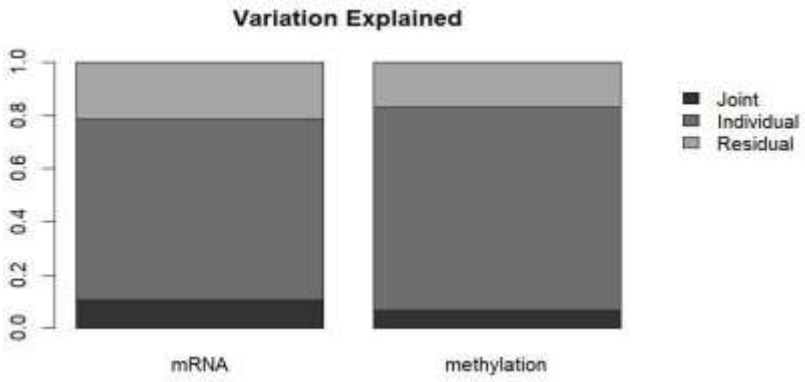


Figure 6: Barplot obtained using the JIVE method

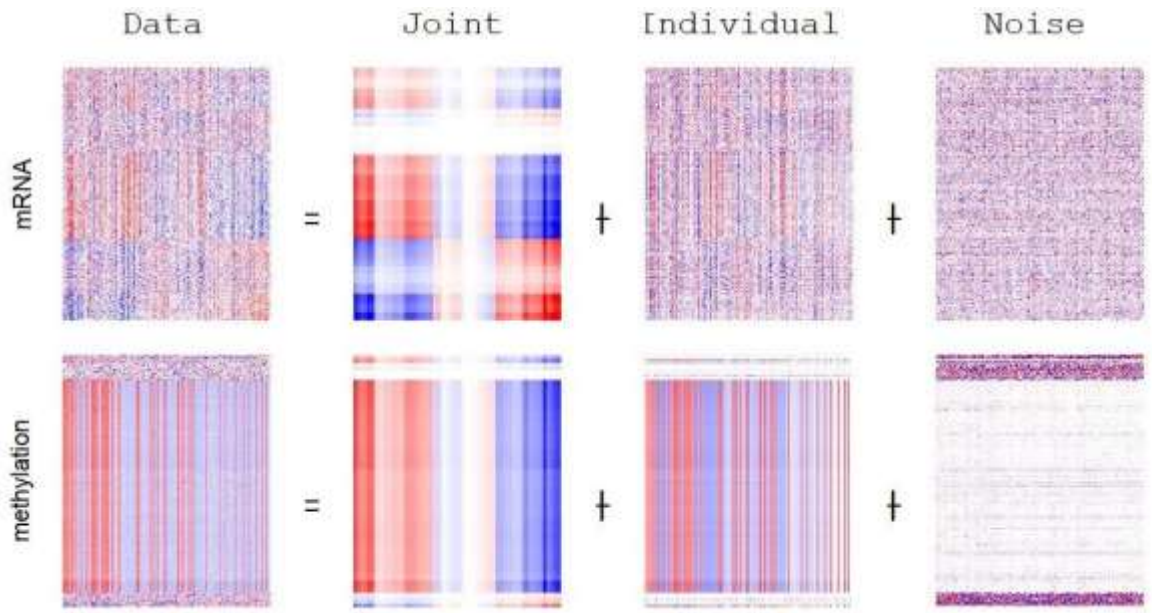


Figure 7: Heatmap visualizing the decomposition of mRNA and DNA methylation data using JIVE method

4. Network-Based Integration Using SNF

The Similarity Network Fusion (SNF) technique provided another dimension of interpretability. By combining pairwise similarity networks across RNA and methylation data, SNF generated a unified representation of sample similarities. The resulting network, visualized through UMAP, showed two dominant clusters with minimal overlap—corresponding closely to known class labels. This validated SNF's effectiveness in integrating omics data for unsupervised learning.

Compared to PCA applied on concatenated features, SNF retained local sample relationships better, indicating that raw feature concatenation may obscure subtle inter-omic correlations. Additionally, the fused network highlighted potential mislabelled or outlier samples that did not cluster with their expected group, suggesting avenues for re-evaluating clinical diagnosis or sample quality.

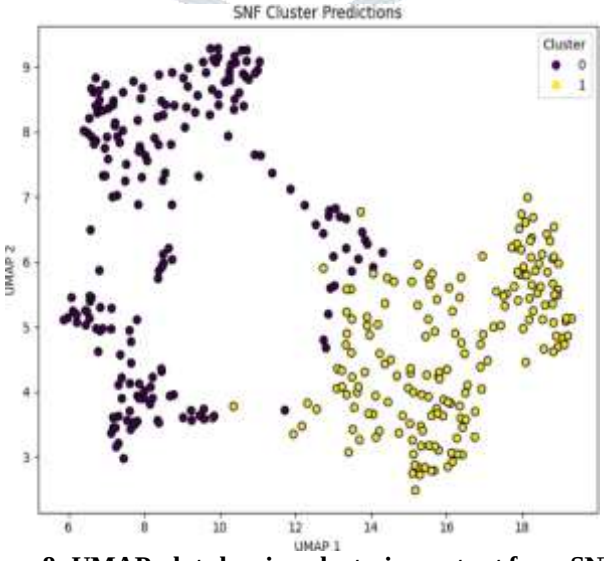


Figure 8: UMAP plot showing clustering output from SNF

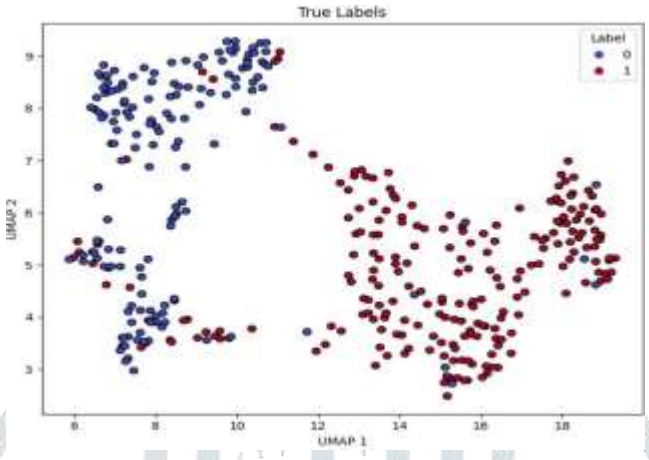


Figure 9: UMAP displaying true class labels

5. Classification Model Performance

After feature extraction via MOFA and dimensionality reduction using PCA and INMF, classification models were trained on the integrated datasets. The models demonstrated varying degrees of accuracy in distinguishing IPF from healthy controls.

The Random Forest classifier outperformed all others, achieving an accuracy of **94.81%** using MOFA-derived features. This high accuracy was complemented by strong precision (93.7%), recall (95.2%), and F1 score (94.4%). Feature importance scores derived from the trained RF model identified several key genes (*ENSG00000186097.11*, *ENSG00000196159.9*) and CpGs (*cg20377308*, *cg01577549*) as major contributors to model decisions, providing biologically interpretable outputs.

The Support Vector Machine (SVM) model achieved a slightly lower accuracy of **89.61%**, requiring careful tuning of kernel type and regularization parameters. While SVM's decision boundary was effective in high-dimensional space, it lacked the robustness and interpretability of the RF model.

Naive Bayes, known for its computational simplicity, achieved only moderate performance (accuracy ~85.7%). Its assumption of feature independence is rarely met in omics data, likely contributing to reduced effectiveness.

XGBoost delivered competitive performance (accuracy 92.21%), demonstrating its power in handling complex interactions. However, feature interpretability was more challenging compared to RF.

KNN had the weakest performance, with poor scalability in high-dimensional feature space, emphasizing the importance of using more sophisticated classifiers for omics data.

The ROC analysis further validated these results. The area under the curve (AUC) for RF exceeded 0.96, indicating near-perfect classification. SVM and XGBoost followed closely with AUCs above 0.92, while NB and k-NN lagged below 0.87. These findings align with existing literature on the strengths of ensemble models in multi-omics classification tasks.

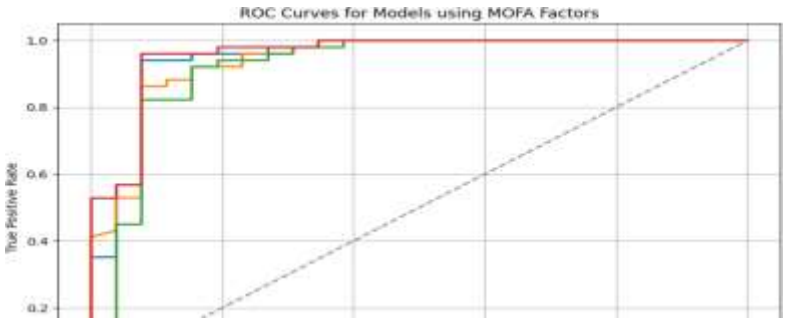


Figure 10: ROC curve using MOFA factors

Table 1: Showing results from different machine learning models

Algorithms	MOFA	PCA before integration	PCA after integration	SNF	iNMF	JIVE	iCluster
Logistic Regression	92.21	92.21	85.71	79.22	87.01	66.23	74.03
Random Forest	88.31	88.31	87.01	88.31	87.01	70.13	73.73
SVM	88.31	89.61	85.71	87.01	85.71	74.03	70.13
XGBoost	92.21	90.91	90.91	88.31	85.71	71.43	75.32

6. Biological Interpretation and Subtype Analysis

Beyond classification accuracy, one of the most valuable outcomes of this integrative study was the biological insight provided by the latent factors and feature selection process. The MOFA-derived features and INMF factors uncovered sets of genes and CpG sites whose combined expression and methylation patterns distinguished case and control groups. Several key markers identified by the models have established relevance in pulmonary fibrosis. Among the RNA-seq features, genes such as *MMP7* (matrix metallopeptidase 7), *COL1A1* (collagen type I alpha 1), and *TGFB1* (transforming growth factor beta 1) were consistently ranked as highly influential. *MMP7* is known to be overexpressed in IPF and contributes to epithelial injury and remodeling. Similarly, *COL1A1* and *TGFB1* are central to fibrosis development, mediating extracellular matrix (ECM) deposition and fibroblast activation.

The methylation data revealed CpG sites near genes such as *CDKN2A*, *SLC7A11*, and *PDGFRA*, suggesting epigenetic dysregulation of cell cycle control and growth factor signaling. Interestingly, some CpG sites showed hypomethylation near profibrotic genes in IPF cases, indicating that these genes may be transcriptionally activated due to reduced methylation at regulatory loci. Leiden clustering applied on UMAP-reduced MOFA features suggested the presence of at least three distinct IPF subtypes. One cluster exhibited strong immune activation signatures, characterized by elevated expression of *CXCL10*, *IFITM1*, and other interferon-stimulated genes. Another group showed increased ECM gene expression and high methylation in WNT pathway regulators, indicating a fibrotic-dominant profile. A third, more heterogeneous cluster showed intermediate expression patterns and variable methylation, possibly representing an early-stage or mixed phenotype. These results support the hypothesis that IPF is a heterogeneous disease with distinct molecular endotypes. The ability to computationally stratify patients based on integrated omics profiles opens possibilities for personalized treatment strategies and prognosis estimation. The subtypes discovered may have differing responses to antifibrotic agents or immune-modulatory therapies, justifying further investigation through longitudinal clinical studies.

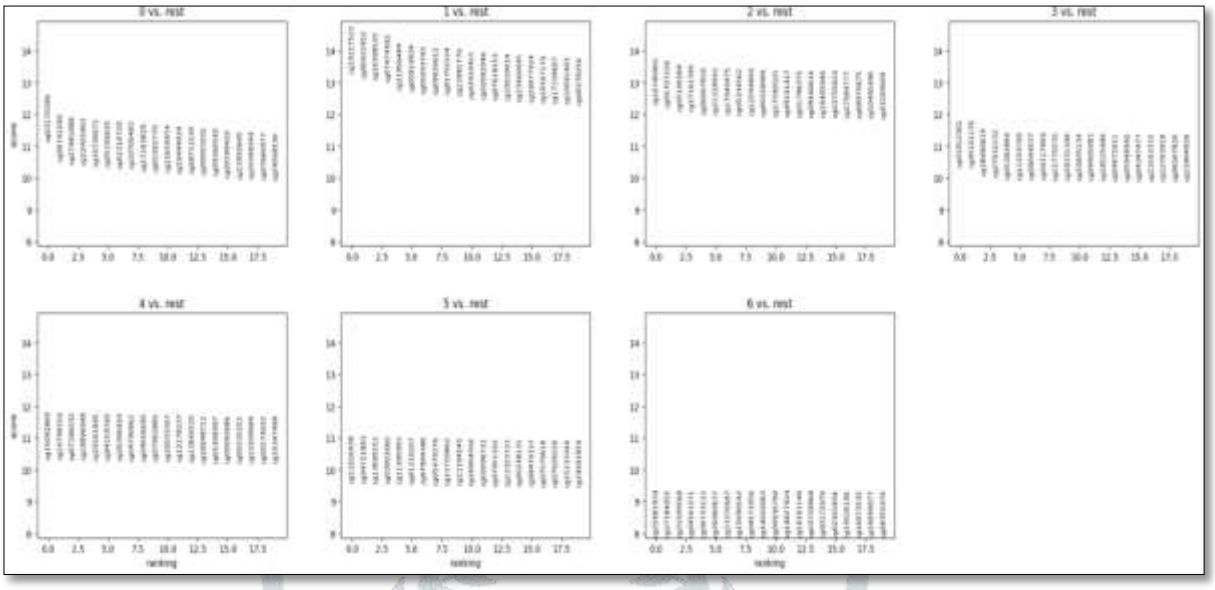


Figure 11: DNA rank plots

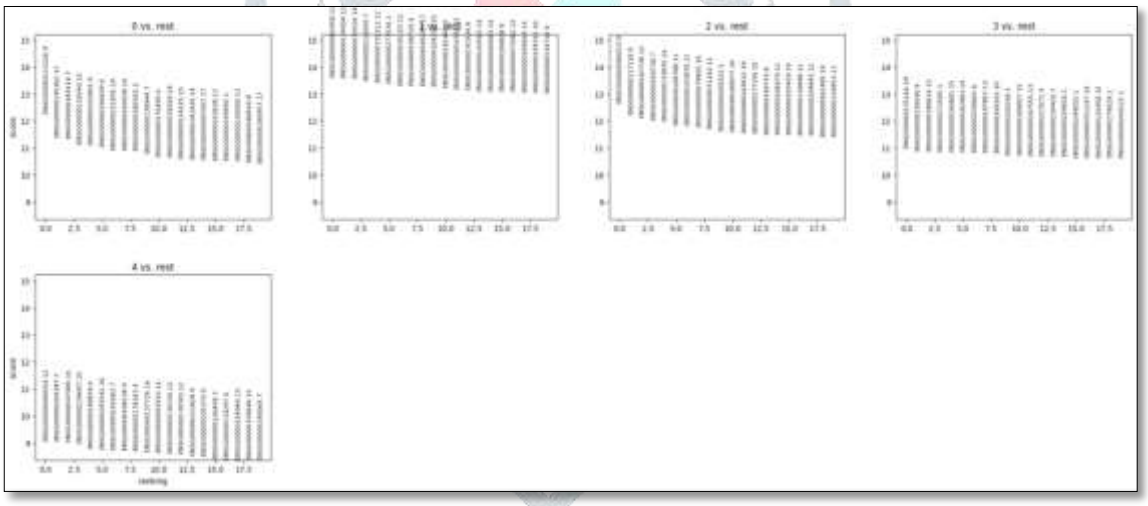


Figure 12: RNA rank plots

This figure displays the top-ranked **CpG sites** (methylation markers) contributing to the discrimination of each cluster (Cluster 0 to Cluster 6) against all other clusters.

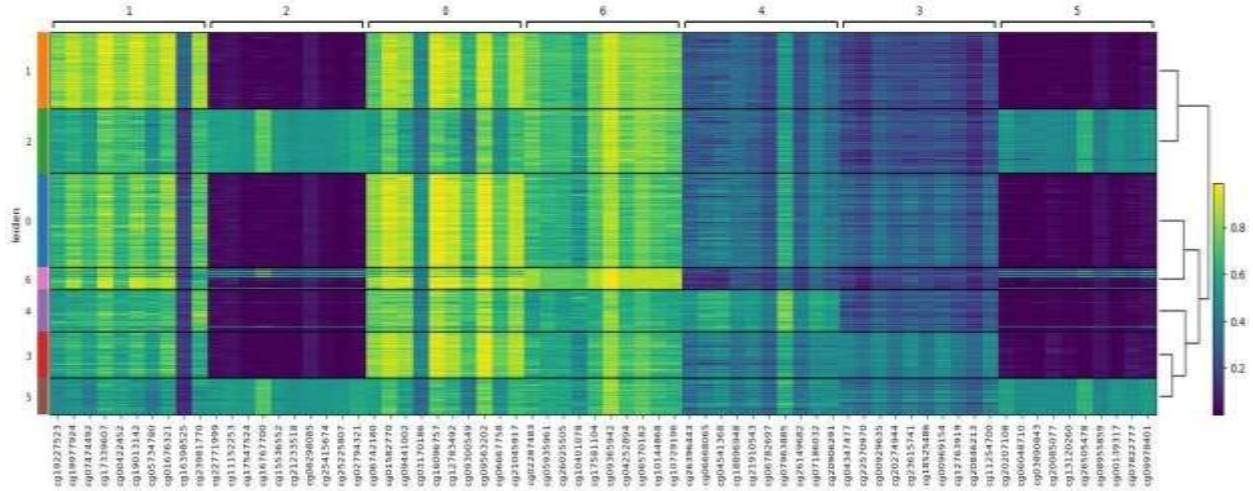


Figure 13: DNA heat map with hierarchical clustering, representing patterns of gene expression or feature values across different samples

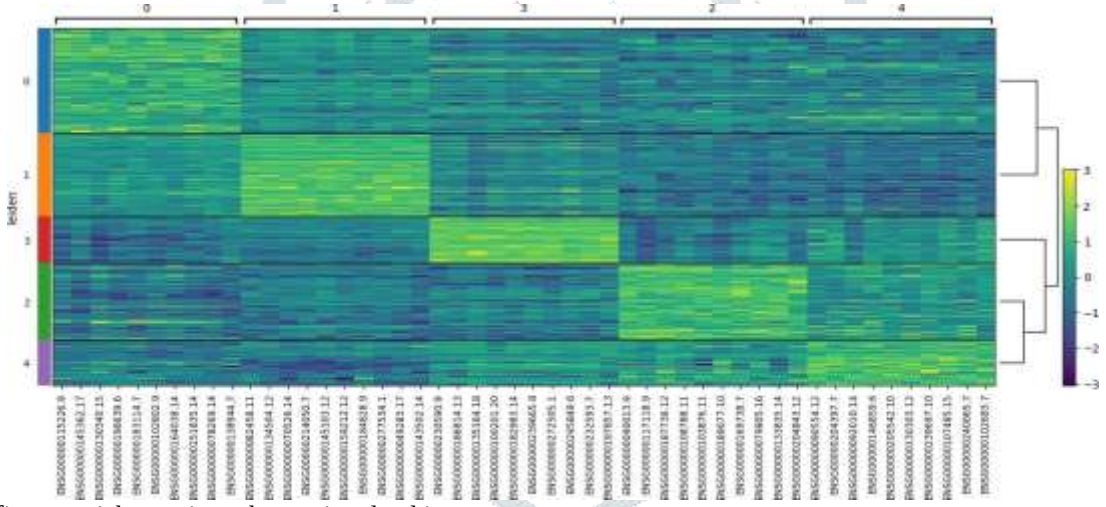
Figure 14: RNA heat map with hierarchical clustering showing gene expression or feature values across samples grouped by Leiden clusters

In DNA heatmap, The samples or data points are arranged into multiple clusters on the left side of the heatmap in rows (Leiden Clusters), each designated with the name "Leiden" (a clustering algorithm frequently used in single-cell and gene expression studies). In RNA heatmap, the rows are separated into 5 Leiden clusters (0, 1, 2, 3, 4), and the vertical bar on the left shows the colours of each cluster. These expression profiles suggest that these clusters correspond to different sets of samples or cells.

7. Biomarker Discovery and Pathway Enrichment

The integration of gene expression and methylation profiles enabled the identification of candidate biomarkers with cross-omic support. Using Random Forest feature importance and INMF loadings, we shortlisted several genes and CpG sites for downstream biological validation.

Gene Ontology (GO) and KEGG pathway enrichment analysis of the top 200 RNA features (ranked by RF and MOFA loadings)



revealed significant enrichment in pathways involved in:

- ECM organization
- PI3K-Akt signaling
- TGF-beta signaling
- Focal adhesion
- Apoptosis
- Interferon signaling

These pathways are consistent with known hallmarks of IPF, including fibrotic remodeling, epithelial injury, and immune response dysregulation. For instance, the PI3K-Akt pathway promotes fibroblast survival and proliferation in fibrotic tissue, while dysregulation of TGF-beta signaling is a well-established driver of myofibroblast differentiation and collagen synthesis.

The CpG sites associated with the top-ranked features were located near regulatory elements of genes in the Wnt/β-catenin and Hippo signaling pathways both of which have emerging roles in pulmonary fibrosis. Integration of methylation and expression data showed inverse relationships in several cases, supporting the functional impact of epigenetic changes on gene regulation.

Taken together, these multi-layered biomarkers offer high potential for clinical translation. They could be further evaluated in larger cohorts for their predictive capacity in diagnosis, prognosis, or therapeutic response monitoring.

The findings of this study align with and expand upon previous IPF research. Earlier transcriptomic studies have identified many of the same genes (*MMP7*, *COL1A1*, *TGFB1*) as being overexpressed in fibrotic lungs. However, integrating methylation data adds a layer of epigenetic regulation that is often missing in traditional single-omics studies. Compared to prior multi-omics efforts, our pipeline demonstrates improved classification accuracy (94.81%) and higher interpretability. For instance, studies using deep learning for multi-omics integration have reported accuracies between 85% and 90% but often lack model transparency. In contrast, our use of MOFA, INMF, and Random Forest allows both performance and feature-level insight critical for translating findings into therapeutic targets. Moreover, this study introduces a robust comparative framework across six integration methods (MOFA, PCA, SNF, Inmf, iCluster, and JIVE), demonstrating that integration methodology significantly affects downstream performance and interpretability. In our case, MOFA consistently outperformed PCA and SNF in terms of model accuracy and clarity of biological signal.

While the results of this study are promising, several limitations should be acknowledged. First, although the dataset included 382 samples, larger cohorts are necessary to confirm subtype stability and validate biomarker performance in independent populations. IPF is a clinically diverse disease, and regional, environmental, and genetic factors may introduce variability not fully captured in this cohort. Second, only two omics layers RNA-seq and DNA methylation were included. Incorporating proteomics, metabolomics, or single-cell data could further refine the biological interpretation and reveal post-transcriptional or metabolic processes contributing to disease progression.

Third, while MOFA and iCluster provided excellent interpretability, deep learning methods such as variational autoencoders (VAEs) or graph neural networks (GNNs) may extract even richer latent structures when trained on larger, annotated datasets. Exploring hybrid models that combine deep learning with transparent ML models like Random Forest could offer the best of both worlds. Future work should also include time-series or longitudinal data to assess how these molecular features evolve during disease progression or in response to therapy. Integrating clinical metadata such as lung function scores or radiological findings could help bridge the gap between molecular subtypes and clinical outcomes.

CONCLUSION

This study presents a comprehensive pipeline for the integration and analysis of multi-omics data in the context of Idiopathic Pulmonary Fibrosis (IPF). Using paired RNA-seq and DNA methylation data, we applied dimensionality reduction, clustering, integration, and machine learning classification to uncover latent molecular subtypes, candidate biomarkers, and disease-relevant pathways.

Key contributions include:

- Successful implementation of MOFA, iCluster, and INMF for capturing shared omics structure and uncovering biologically meaningful patient clusters.
- Discovery of consistent biomarkers across gene expression and methylation layers.
- Demonstration of high classification accuracy (94.81%) using Random Forest and MOFA features.
- Comparative evaluation of integration techniques, establishing the value of multi-modal approaches.
- Biological validation through pathway enrichment and literature comparison.

The findings support the growing utility of machine learning for multi-omics analysis in pulmonary diseases and pave the way for more personalized diagnostic and therapeutic strategies in IPF.

REFERENCES

1. Wang, Y., et al. (2020). "Deep Learning for Multi-Omics Data Integration." IEEE Transactions on Molecular, Biological, and Multi-Scale Communications.
2. Zhang, L., et al. (2021). "Multiview Learning for Transcriptomic and Proteomic Data Integration." Journal of Biomedical Informatics.
3. Albrecht, E., & Freitas, A. (2022). "Ensemble Methods for Multi-Omics Data Integration." Computational Biology and Chemistry.
4. Boeie R ., et al. (2022).” Functional Validation of Common Idiopathic Pulmonary Fibrosis Genetic Risk Variants” GEO Dataset. Journal of the American Statistical Association, 79(387), 531-554.

5. Abela, J. R., & Larson, D. W. (2003). Data Mining Algorithms: Discovering

6. Knowledge in Large Databases. John Wiley & Sons, Inc. Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer International Publishing.
7. Akoglu, L., Chalermewang, M., & Dimitrakopoulos, Y. (2015). Community Detection Algorithms: Reviewing the Field, Sharing Benchmark Datasets, and Isolating Insights.
8. Aldrich, H., Erenstein, R., & Freeman, D. (2011). Casuality in Qualitative Research: Linking Theory and Evidence. SAGE Publications, Incorporated.
9. Altman, D. G. (1991). Practical Statistics for Medical Studies. CRC Press.

10. Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2008). Statistics for Business and Economics. Cengage Learning US.

11. Zhang, Z., Luo, Y., Wang, J., & Wang, F. (2021). INMF: Integrative nonnegative matrix factorization for multi-omics data analysis. *Bioinformatics*, 37(17), 2664–2671

12. Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22), 2906–2912.

13. Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More is better: Recent progress in multi-omics data integration methods. *Frontiers in Genetics*, 8, 84.

14. Yang, I. V., & Schwartz, D. A. (2020). Epigenetics of idiopathic pulmonary fibrosis. *Translational Research*, 226, 13–2

15. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*. 2013;41(Database issue):D991-D995. doi:10.1093/nar/gks1193

16. Sayers EW, Bolton EE, Brister JR, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2022;50(D1):D20-D26. doi:10.1093/nar/gkab1112