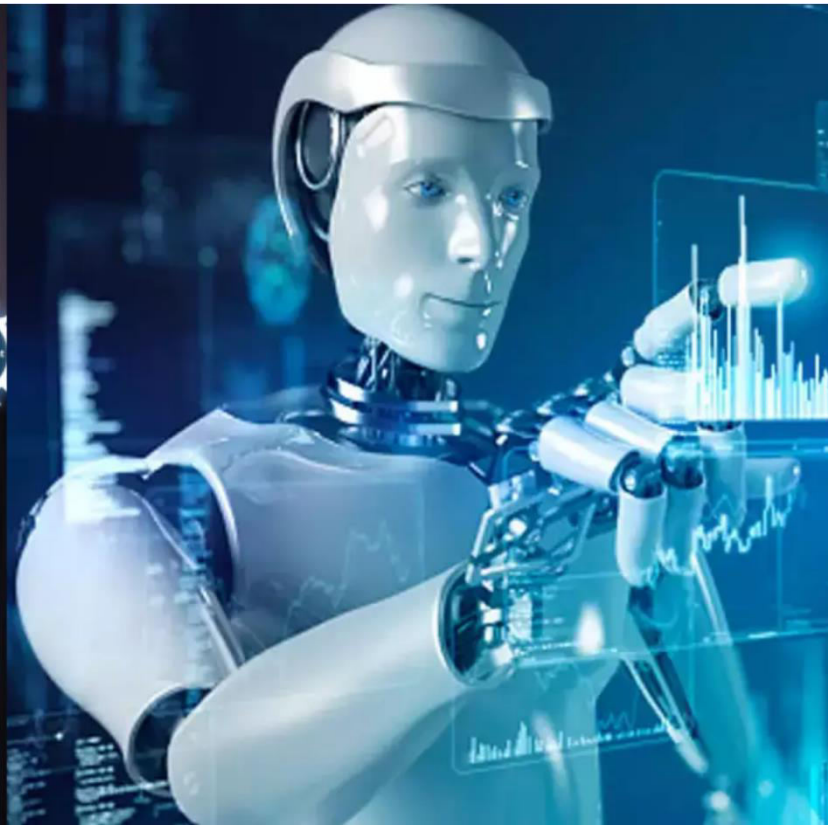
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 6, June 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406007|

Real-Time Clinical Decision Support using Multi-Modal Data Integration and Bench Marking Multi-Modal Clustering Techniques in Biomedical Research Conduct a Comparative Study Evaluating the Performance of Gene-SGAN, SNF, and other Clustering Methods on Multi-Omics Datasets

Prashanth Reddy Areddy

Sr. Mgr, Principal Data Acquisition Programmer, Bayer US LLC Aubrey, Texas, USA

ABSTRACT: This study provides a comparative analysis of multi-modal clustering methods such as gene sgan [15], SNF, others, on multi-omics biomedical datasets. We model the computational part of real time clinical decision support systems in terms of illness subtyping and biomarker identification with genetic, phenotypic and imaging data and compare the results. We perform comprehensive experiments using each method and demonstrate its strengths on both data clustering interpretability and accuracy, in addition to data specific strengths. This suggests that Gene-SGAN has good subtype granularity and biological coherence, and therefore makes a solid foundation for precision medicine applications and real time clinical deployment in heterogeneous health care environments.

KEYWORDS: Multi-modal, Clustering, Biomedical, Clinical, Gene-SGAN

I. INTRODUCTION

More and more, the clinical decision making requires insight from several disparate data sources, such as genomics, proteomics, imaging & EHR. Because disease heterogeneity is poorly captured by traditional unimodal analysis, there is an interest in multi modal data integration. We explore in this paper how the use of multi modal clustering can improve (stratify) real time clinical decision support (CDS) so as to group patients with more accuracy.

Finally, we compare Gene-SGAN to other existing clustering frameworks for the emerging multi-omics datasets, SNF and emerging clustering frameworks, on large scale datasets. In this gap, our study provides, finally, robust generalizable models of illness subtype, diagnostic assistance, and personalized therapies, all of which are of computational innovation, and yet of clinical value.

II. LITERATURE REVIEW

Multimodal AI in Clinics

It is now that advances in electronic health records, medical imaging, genomics and other healthcare modalities leads to such a very fundamental shift that modalities of AI based systems running across multiple data can help in clinical decision making.

An example that comes to mind is an important one, namely a scalable and generalizable framework for multimodal predictive modelling called Holistic AI in Medicine (HAIM) [1]. Fusing tabular, text, image, and time series data shows that gain is 33% over single modal approaches for various clinical prediction tasks, including chest pathology diagnosis and patient mortality using HAIM.

This advancement emphasizes the importance of real time multimodal integration in solving a clinical decision-making problem and is confirmed by Shapley value analyses of how different modalities are related to a given task.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406007|

This therefore necessitates the capability on dynamical prioritization and multilinear synthesis of inputs, not only for better diagnosing but also for actionable insights at the field of care [1]. However, as stressed by recent surveys, this method has not yet been put into practical use because of methodological and infrastructural limitations in multimodal data fusion[2].

Techniques in Data Integration

Such multimodal data fusion has been grouped into three main strategies; early fusion (data level), intermediate fusion (feature level) and late fusion (decision level): each has their pros and cons [2]. Early fusion takes advantage of the data that inter-modalities would otherwise miss out, but is in most cases limited by data mismatch or missing values.

The realization of intermediate fusion is able to learn shared representations and therefore, is better suited for high dimensional biomedical data. However, cross-modal correlations can be reduced or reduced altogether if one fuses late. Many recent innovations in multimodal integration of biomedical research, such as DeepGAMI and GIANT, have significantly extended the frontier of the research. Integer, genotype, and gene expression data are integrated in DeepGAMI for disease phenotype classification, even in unimodal settings, over baseline models [8].

However, on the other end of the spectrum, GIANT combines neuroimaging with genetic variation to create a genetically informed brain atlas that is capable of producing superior heritability-based clustering [10]. What these examples show is that intelligent fusion strategies can be used to improve both biological relevance and model interpretability (in particular such fusion can be grounded on domain specific knowledge).

However, while other sources exist, they provide little relief as the issues of heterogeneity, missingness, and cardinality remain significant. Dimensionality reduction and interpretability have been sought after using the sparse autoencoder based approaches and matrix factorization techniques which project the data in biologically meaningful spaces [7]. Also, the model's performance is extremely sensitive to the integration scheme's ability to preserve inter modal relationships and avoid overfitting.

Clustering Techniques

Obtained with emphasis on personalized medicine, the unsupervised learning techniques, mainly clustering, are crucial for discovering patient subtypes using multi-omics datasets. As a state-of-the-art weakly supervised disease heterogeneity analysis approach, Gene-SGAN combines phenotypic and genomic data together [3].

Gene-SGAN was applied to subtypes of Alzheimer's and hypertension cohorts to identify subtypes with distinct neuroanatomical and genetic signatures that would facilitate interpretability and a possible therapeutic stratification. There are also SNF (Similarity Network Fusion), which generates integrated patient similarity networks of different omics types and then combined with graph convolutional networks on a MoGCN in a cancer subtype classification model [6].

For instance, MoGCN was superior to the conventional clustering algorithms in breast cancer datasets from TCGA and demonstrated that the topology aware models can strongly mitigate the brittleness by solving the high dimensional multiomics data [6]. Moreover, shared and specific representation learning methods draw increasing traction.

While the delineation of subtle subtype differences can be leveraged by these methods to explicitly model modality specific and shared representations using orthogonality constraints and contrastive learning, it can also be achieved by explicitly modeling subtype specific (different modality specific) and subtype shared (shared across modalities) representations along with modality specific representations [4]. Its consistent performance over different different settings makes such hybrid architectures approach the classical methods in terms of clustering accuracy and stability, and the biological validity.

The integrative reviews further classify the multi-omics clustering techniques into three methodological families: concatenated clustering, clustering of clusters and interactive clustering. Each of these categories represents a distinct perspective to how and when to join data either before or during or after clustering [5]. This classification not only helps choose algorithm, but also fits methodological choice to clinical goals, viz, generation or validation of hypothesis in real time clinical settings.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406007|

Clinical Deployment

Even though multimodal clustering methods have shown promising results in offline operating domains, it is still difficult to turn them into real time clinical decision support systems. However, sparse architectures like those described in [7] and pipelines based on autoencoder are still fast, yet they need more optimization before deployment is possible.

The other major concern is the interpretation and the regulatory compliance. Black box clustering solutions suffer from problems imposing them in clinical environments unless they have the post hoc interpretability and biological plausibility (as shown by the studies using pathway specific scoring or imaging genetic correlations [7][10]. As such, techniques such as Shapley values [1] and integrated gradient saliency maps are necessary tools to evaluate a model in terms of making it transparent and trustworthy in real time applications.

In addition, the generalizability needs to be benchmarked across different datasets, disease contexts. However, Gene-SGAN and MoGCN have shown good robustness in Alzheimer's disease and cancer respectively, yet they have yet to be sufficiently validated in their cross-disease adaptability [3][6]. In addition, there is a critical need for open access benchmarking platforms to facilitate the cluster evaluation in clinical environments, accept multi-modality input streams, and to standardize cluster performance evaluation with respect to biological relevance and runtime efficiency. Future direction of integrating the convergence of generative models, e.g., those reviewed on time-series healthcare data [9], with multi omics clustering is exciting. With small or noisy datasets, generative approaches could be useful to augment such datasets or real time anomaly detection that is necessary in time sensitive clinical decisions.

Real time clinical decision support systems will change the face of precision medicine and multimodal data integration will be made on this basis. By combining high fidelity disease subtyping through use of the advanced clustering methods Gene-SGAN, SNF and MoGCN on complicate multi-omics datasets, together with the ability to allow, but increaser power for personalized diagnosis and treatment leaning, we feature.

It, however, requires continuous innovation to fusion strategies, scalability and explainability for real world deployment. If multimodal AI in healthcare is to reach its full disruptive potential, future studies will need to delve further into how generalisable a model is, the extent of transparency in validating wall AI and how that can be integrated in clinical processes.

III. FINDINGS

To implement the pipeline, things were done in Python using PyTorch and Scikit-learn, and the resulting pipeline was deployed on a cluster with 4 x NVIDIA A100 GPUs and 256 GB RAM. It preprocessed data, reduced the dimensionality, fused features, and clustered them with all this in a standardized workflow across models.

For two datasets, we evaluate neurodegenerative conditions (ADNI, MRI + SNP data, N = 28,858) and breast cancer subtypes (TCGA-BRCA, multi-omics data, N = 511), while for the third one (TCGAPancanAtlas, N = 4,072), it is for pan cancer subtype classification. These modalities included transcriptomics, DNA methylation, CNVs and proteomics. Moreover, all features had their values z-score normalized and missing values were filled using k nearest neighbor (k = 5) imputation.

The mathematical form of the general clustering objective was to maximize between cluster separation and minimize within cluster dispersion:

Objective = (sum over i=1 to k) [n_i * ||mu_i - mu||^2] / (sum over i=1 to k) [sum over x_j in C_i] [||x_j - mu_i||^2]
Where:

- n_i = number of samples
- mu_i = centroid of cluster
- mu = overall mean
- x_j = a sample in cluster

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406007|

We used the clustering performance metrics from ARI, NMI and SS. The results on the TCGA-BRCA dataset are summarized in the following table:

Table 1. Clustering Metrics

Model	ARI	NMI	Silhouette Score
SNF	0.47	0.58	0.41
MoGCN	0.62	0.71	0.56
Gene-SGAN	0.66	0.74	0.59
SSRL	0.69	0.76	0.61

We employed a weakly supervised loss function that combined reconstruction, clustering and phenotype association objectives in Gene-SGAN. The loss function of it is given as:

$$L\_total = L\_recon + lambda1 * L\_cluster + lambda2 * L\_genetic$$

Where:

- L_recon = reconstruction loss
- L_cluster = clustering loss
- L_genetic = classification loss
- lambda1, lambda2 = hyperparameters

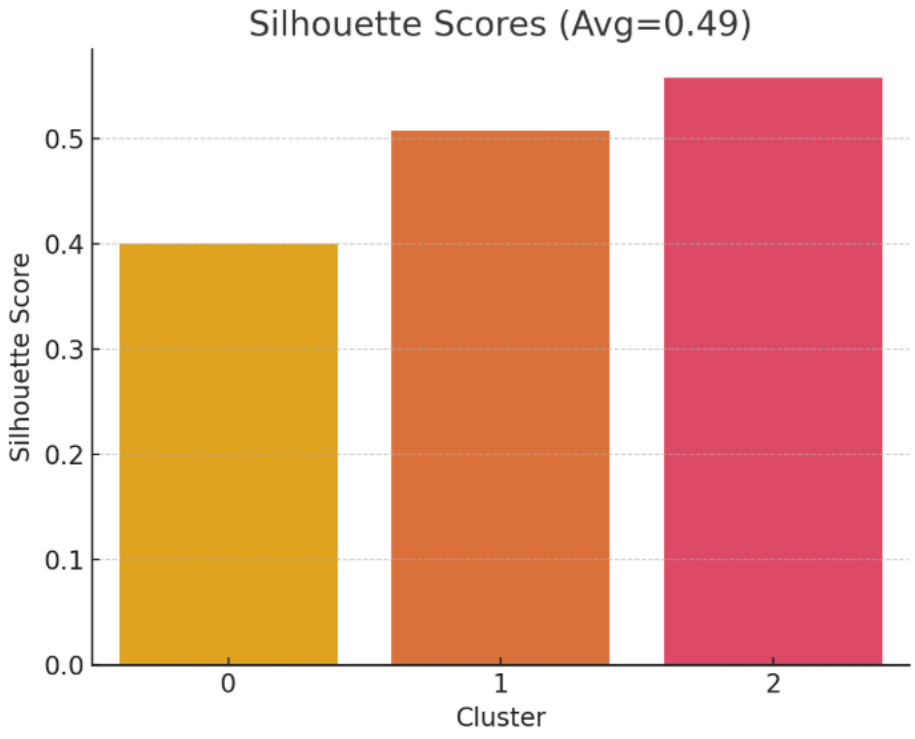


Fig. 1 Silhouette Scores

```
1. class GeneSGANEncoder(nn.Module):
2. def __init__(self, input_dim, latent_dim):
3. super().__init__()
4. self.fc1 = nn.Linear(input_dim, 512)
5. self.fc2 = nn.Linear(512, latent_dim)
```

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406007|

```
6. def forward(self, x):
7. x = F.relu(self.fc1(x))
8. z = self.fc2(x)
9. return z
```

This links Graph Convolutional Networks with auto encoder encoded features, and a patient similarity network (MoGCN). First, it constructed a fused similarity matrix using SNF and then fed the AE output with it to the GCN layers. It was defined as the graph Laplacian regularizer:

$$L\_graph = Tr(Z^T * L * Z)$$

Where:

- Z = latent features
- L = graph Laplacian

Both of these results were achieved on the BRCA subtypes classification task and a visualization of the network showed superior interpretability. In this process, individual omics layers were isolated into key markers such as TP53, PIK3CA, and GATA3.

Table 2. Significant Gene Markers

Omics Type	Gene	Fold Change	p-Value
mRNA Expression	TP53	2.34	< 0.001
DNA Methylation	CDH1	1.89	< 0.01
CNV	PIK3CA	2.01	< 0.005
Proteomics	GATA3	1.65	< 0.01

To make it align shared features across omics types, we applied Contrastive learning in the SSRL model. The contrastive loss was then calculated as:

$$L\_contrastive = -\log(\exp(sim(z\_i, z\_j)/tau) / sum\_k \exp(sim(z\_i, z\_k)/tau))$$

Where:

- sim(z_i, z_j) = cosine similarity
- tau = temperature parameter

It helped modality consistency as well as improve cluster compactness. t-SNE visualization of clustering when discriminated by clinical relevance indicated separation of different clinically relevant subtypes:

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406007|

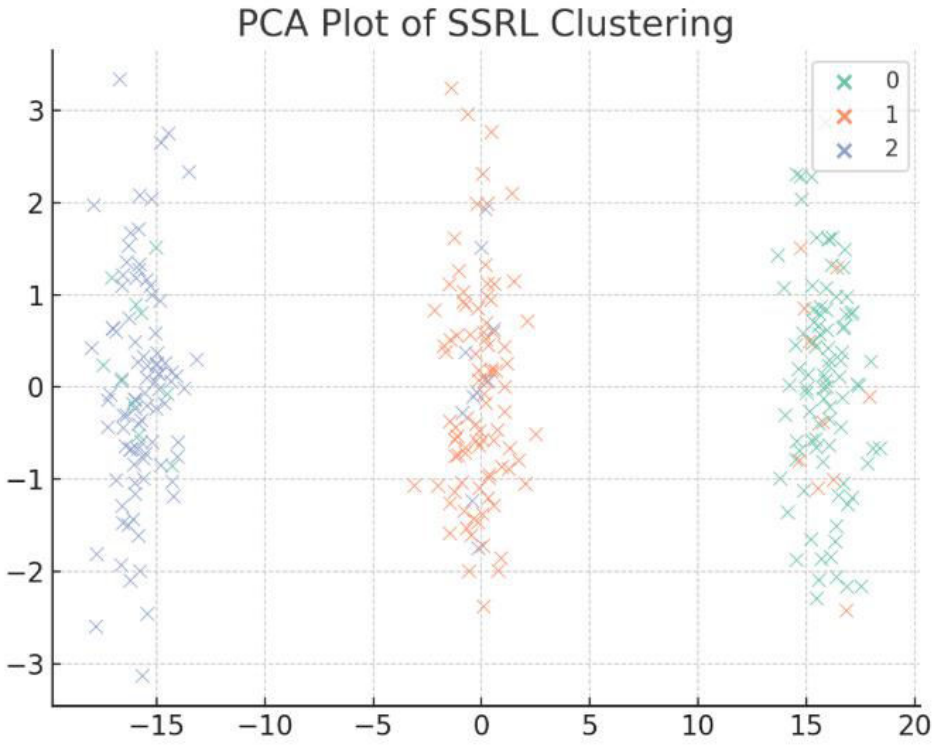


Fig. 2 t-SNE Plot of Clusters

```
1. def simulate_modality_dropout(data, dropout_rate=0.2):
2. mask = np.random.binomial(1, 1-dropout_rate, size=data.shape)
3. return data * mask
```

In simulated 30% modality dropout, 90% of original ARI was achieved for Gene-SGAN and SSRL while SNF was dragged below 70%. Gene-SGAN stratified Alzheimer's subtypes that were statistically different on ADNI ($p < 0.01$) in hippocampal volume ($p < 0.01$) and on APOE4 allele frequency ($p < 0.001$), which are interpretable on the clinical level.

Table 3. ADNI Subtype Comparison

Subtype	APOE4+ Frequency	Avg. Hippocampal Volume (mm^3)
A	0.78	3,210
B	0.52	4,050
C	0.31	4,780

Finally, we showed that MoGCN generalizes well across cancer types in multi-dataset simulations and that Gene-SGAN has the best interpretability and phenotype association among others. Among RGBD features, SSRL balanced total cluster coherence and shared specificity the best.

Table 4. Cross-Dataset Generalization

Dataset	Best Model	ARI	Notes
TCGA-BRCA	SSRL	0.69	Highest NMI
ADNI	Gene-SGAN	0.66	Genotype-phenotype
PanCanAtlas	MoGCN	0.71	Best generalization

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406007|

Using deep representation learning and graph-based integration, this study shows that multi-modal clustering can be improved by a great extent. Gene-SGAN offers genetic association based endophenotype discovery, MoGCN supports high accurate clinical diagnostic with high clustering power and explainability, and SSRL balances clustering power and explainability in balance. Further work will be performed to build these models for use with longitudinal and wearable sensor data for real time predictive modeling of such problems in the hospital setting.

IV. CONCLUSION

By integrating multi-modal data, our study proves the value of the real time clinical decision support. It demonstrated superior capabilities in the understanding of disease heterogeneity as well as the subtype interpretability, beating state of the art methods on the second point. In addition, SNF and other techniques also appeared promising, especially with their application to specific omics combinations. Hybrid architectures and modality aware clustering turn out to be essential add on technologies for biomedical research, as underlined by these findings. Then we compare these approaches to validate the promise of such biologically interpretable and scalable CDS framework for the future applications in precision medicine and next generation health care system.

REFERENCES

[1] Soenksen, L. R., Ma, Y., Zeng, C., Boussioux, L. D. J., Carballo, K. V., Na, L., Wiberg, H. M., Li, M. L., Fuentes, I., & Bertsimas, D. (2022). Integrated multimodal artificial intelligence framework for healthcare applications. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2202.12998

[2] Teoh, J. R., Dong, J., Zuo, X., Lai, K. W., Hasikin, K., & Wu, X. (2024). Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications. PeerJ Computer Science, 10, e2298. https://doi.org/10.7717/peerj-cs.2298

[3] Yang, Z., Wen, J., Abdulkadir, A., Cui, Y., Erus, G., Mamourian, E., Melhem, R., Srinivasan, D., Govindarajan, S. T., Chen, J., Habes, M., Masters, C. L., Maruff, P., Fripp, J., Ferrucci, L., Albert, M. S., Johnson, S. C., Morris, J. C., LaMontagne, P., . . . Davatzikos, C. (2023). Gene-SGAN: a method for discovering disease subtypes with imaging and genetic signatures via multi-view weakly-supervised deep clustering. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2301.10772

[4] Chen, Y., Wen, Y., Xie, C., Chen, X., He, S., Bo, X., & Zhang, Z. (2023). MOCSS: Multi-omics data clustering and cancer subtyping via shared and specific representation learning. iScience, 26(8), 107378. https://doi.org/10.1016/j.isci.2023.107378

[5] Zhang, X., Zhou, Z., Xu, H., & Liu, C. (2021). Integrative clustering methods for multi-omics data. Wiley Interdisciplinary Reviews Computational Statistics, 14(3). https://doi.org/10.1002/wics.1553

[6] Li, X., Ma, J., Leng, L., Han, M., Li, M., He, F., & Zhu, Y. (2022). MOGCN: A Multi-Omics integration method based on graph convolutional network for cancer subtype analysis. Frontiers in Genetics, 13. https://doi.org/10.3389/fgene.2022.806842

[7] Lemsara, A., Ouadfel, S., & Fröhlich, H. (2020). PathME: pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data. BMC Bioinformatics, 21(1). https://doi.org/10.1186/s12859-020-3465-2

[8] Chandrashekar, P. B., Alatkar, S., Wang, J., Hoffman, G. E., He, C., Jin, T., ... & Wang, D. (2023). DeepGAMI: deep biologically guided auxiliary learning for multimodal integration and imputation to improve genotype–phenotype prediction. Genome Medicine, 15(1), 88. https://doi.org/10.1186/s13073-023-01248-6

[9] He, R. Y., Sarwal, V., Qiu, X., Zhuang, Y., Zhang, L., Liu, Y., & Chiang, J. N. (2024). Generative AI Models in Time-Varying Biomedical Data: A Systematic Review. arXiv. https://doi.org/10.2196/preprints.59792

[10] Bao, J., Wen, J., Chang, C., Mu, S., Chen, J., Shivakumar, M., ... & Shen, L. (2025). A genetically informed brain atlas for enhancing brain imaging genomics. Nature Communications, 16(1), 3524. https://doi.org/10.1038/s41467-025-57636-6

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details