GMR The Original

An Advanced Multi-Omics Data Integration Framework Using Machine Learning and Bioinformatics Techniques for Tumor Genomics and Cancer Subtype Identification

Indu Purushothaman [1], Jayannan J [2], Sindhu Subramani [3], Sathasivam Sivamalar [4], Gowthami Priyadharshini [5], Indu Purushothaman [6], Pooraninaga Lakshmi J [7],

[1]Assistant Professor, Department of Research, Meenakshi Academy of Higher Education and Research. indu@maher.ac.in, ORCID: 0000-0002-3996-7927
[2]Department of General Medicine, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research. jayannan@maher.ac.in, ORCID: 0000-0002-0245-7966
[3]Assisant Professor, Department of Oral Pathology, Meenakshi Ammal Dental College and Hospital, Meenakshi Academy of Higher Education and Research. sindhus@maher.ac.in, ORCID: 0009-0005-2550-6408
[4]Scientist, Department of Research, Meenakshi Academy of Higher Education and Research. malarss@maher.ac.in, ORCID: 0000-0002-8774-2773
[5]Lecturer, Meenakshi College of Pharmacy, Meenakshi Academy of Higher Education and Research. gowthamip@maher.ac.in, ORCID: 0009-0008-8568-2210
[6]Department: Conservative Dentistry and Endodontics, Designation: Senior lecturer, College address with pin code: Sree Balaji Dental College and Hospital, (Affiliated to Bharath Institute of Higher Education and Research), Pallikaranai, Chennai – 600100, Email id: pooranijana@gmail.com, Orcid ID: 0000-0001-7507-5283
[7]Assistant Professor, Department of Research, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India- 600087 E-mail: indu@maher.ac.in

ABSTRACT

One of the most significant dilemmas to cancer genomics is tumour heterogeneity, since molecular differences among patients tend to downgrade the efficiency of single-omics analyses in the proper characterization of tumour behaviour and clinical outcome characteristics. Multi-omics data integration gives a global perspective of the intricate regulated interactions of the molecular processes involved in cancer progression and establishment; nonetheless, cross-dimensional assembly of heterogeneous and high-dimensional omics measurements has been a crucial obstacle in computation. In this work, we suggest a highly developed machine learning-based multi-omics information integration framework, which will improve the tumour genomic analysis and will help to make the cancer subtype identification strong. The suggested model uses representation learning, which is based on a deep learning approach, and attempts to combine various omics layers, such as genomic, transcriptomics, and epigenomics data, into a unified latent feature space. The quality of integration is determined in a systematic manner through several quantitative measures such as reconstruction error, values of clustering validity and stability analysis and compared with the standard methods of integration. The integrated representations are then subjected to unsupervised clustering solutions to determine discrete cancer subtypes, which are then performed on supervised classification models to confirm the predictability of the subtypes. Through experimental findings, it is indicative of the fact that the proposed framework has better integration quality and better separations of subtypes as compared with the baseline methods. Moreover, the subtypes identified have high biological and clinical significance as they share a considerable molecular signature and differ largely in the outcomes of patient survival. On the whole, this analysis indicates that a multi-omics integration using machine learning is efficient in the area of tumour genomics and can play a significant role in a more accurate cancer analysis and individualization in the therapy approach.

Key words: *Tumor genomics, cancer subtype identification, machine learning, deep learning, auto encoder, Variational auto encoder.*

INTRODUCTION

Cancer is a disease that is highly heterogeneous, and features multi-faceted molecular changes that are diversified in human beings, among tumours and at different stages of disease progression. This heterogeneity is one of the greatest problems of tumour genomics, with various genetic, epigenetic, and transcriptional mechanisms acting together in the formation of tumour initiation, progression, and response to therapy. Cancer

genomics projects at large scale have produced multi-ome-scale data sets of complementary molecular layers, including genomics, transcriptomics and epigenomics, providing the opportunity to describe tumours more holistically. Nevertheless, it is not a trivial task to extract coherent and biologically informative data out of such heterogeneous sources because the aim is to determine clinically relevant cancer subtypes that are indicative of underlying molecular diversity and patient outcomes (Mo et al., 2013; Ramazzotti et al., 2018).

Traditional cancer analysis methods have mostly been based on single-omics, which does not easily meet the complexity of tumour biology. Even though initial integrative techniques, such as matrix factorization, similarity-based fusion, and joint clustering, have shown to be better than single-omics analyses, they are often not as well positioned to deal with high dimensionality, missing data, and nonlinear interactions between the layers of the omics analyses (Lock and Dunson, 2013; Wang et al., 2014). In addition, the conventional integration paradigm might not maintain biologically significant shared signals and modality-specific signals, which results in the poor subtype separation and diminished interpretability (Nguyen et al., 2017; Rappoport and Shamir, 2019). These constraints explain the necessity of more versatile and dynamic integration strategies with the ability to learn powerful representations with regards to heterogeneous multi-omics data.

The current innovations in machine learning and bioinformatics have dramatically changed the identification of subtypes of cancer since it is now possible to model complex patterns of molecules using data. Autoencoders and graph-based models, which are based on deep learning, have also shown good prospects of learning nonlinear relationships between latent representations among multiple omics modalities (Benkirane et al., 2023; Poirion et al., 2021; Wang et al., 2021). Combined multi-omics analysis has demonstrated the ability to outperform several integrative frameworks, including MOFA, NEMO, MOVICS, and MOGONET in cancer subtyping, prediction of prognosis, and biomarker discovery (Argelaguet et al., 2018; Lu et al., 2020; Rappoport and Shamir, 2019; Wang et al., 2021). However, the current body of research also devotes a little bit of attention to assessing the quality, strength, and resilience of data integration in a systematic manner that is needed to guarantee trustful biological interpretation and clinical usage success (Velten et al., 2022).

Based on these problems, this paper suggests a state-of-the-art machine learning data integration framework to develop tumour genomics and cancer subtype with multi-omics. The most important works of this work are triple. To begin with, we create a scalable framework of heterogeneous omics that is based on deep learning and can learn shared latent representations (Benkirane et al., 2023; Zhang et al., 2020). Second, we perform a full quantitative assessment of the quality of multi-omics integration through a variety of metrics by representation and clustering (Argelaguet et al., 2018; Wang et al., 2014). Third, we show strong cancer subtype discovery with a biological, clinical interpretation, and survival relevance and molecular characterization (Mo et al., 2013; Ramazzotti et al., 2018). This study will not only fill methodological gaps but also biological gaps in developing comprehensive cancer genomics that can be used in the formation of more trusted computational methods to oncology accuracy.

Related Work

The early studies of cancer genomics were mainly on the integration of heterogeneous omics data through conventional statistical methods and matrix factorization-based data integration methods. Common approaches to detect common patterns of molecules were the joint latent variable modelling, Bayesian consensus clustering, and low-rank matrix factorization, which were used extensively across genomics, transcriptomics, and epigenomics data (Lock and Dunson, 2013; Mo et al., 2013). Similarity-based approaches, such as similarity network fusion, also allowed the combination of different omics data types based on building the population-scale networks of patient similarity, which had a better subtype discovery than single-omics analyses (Wang et al., 2014). Although they demonstrated that integrative cancer analysis is possible, such methods were often constrained by linear modelling, prone to noise, and unable to scale to large and complex data (Nguyen et al., 2017; Ramazzotti et al., 2018).

As high-throughput sequencing technologies continue to improve chaotically, machine learning-based approaches to the integration of multi-omics have been afloat in popularity. Unsupervised algorithms like Multi-Omics Factor Analysis and its variants presented probabilistic latent variable based models that could break down shared and modality specific sources of variation in across two or more layers of Omics (Argelaguet et al., 2018; Velten et al., 2022). More recently, those solutions, which are based on deep learning, such as autoencoders and representation learning models, demonstrated better performance at capturing the nonlinear relationships and more complex interactions involved in multi-omics data (Benkirane et al., 2023; Zhang et al., 2020). Integration models based on graphs also boosted the performance of integration of it through the use of

relational structure among patients and molecular features allowing the more accurate classification and discovery of biomarkers (Wang et al., 2021).

Based on representation of integrated features, a range of studies have sought to find cancer subtype by clustering and classification structures. Non-negative matrix factorization, consensus clustering, and network-based clustering are examples of techniques that have extensively been used to identify molecularly distinct cancer subtypes that have proven prognostic value (Lu et al., 2020; Rappoport and Shamir, 2019). Moreover, models of supervised and semi-supervised learning, such as ensemble learning, deep learning, have been used to predict the subtypes of cancer and clinical outcomes using combined multi-omics features (Poirion et al., 2021; Wang et al., 2021). Even though the results of these approaches have been positive in terms of deriving biologically relevant subgroup and enhancing survival prediction, these useful techniques require the quality and strength of the underlying data integration procedure (Ramazzotti et al., 2018).

Nevertheless, in spite of these major progress, there are still a number of limitations and gaps in research in the existing studies of multi-omics integration. Most of the existing strategies focus on predictive performance with little quantitative assessment of integration quality and sample representation robustness (Argelaguet et al., 2018; Wang et al., 2014). In addition, learned latent features are also biologically interpretable, which is especially not the case with deep learning-based models that can be seen as black-box tools (Benkirane et al., 2023; Velten et al., 2022). It is also not necessarily coupled with issues of scalability, omics layer robustness, and generalizability to various types of cancers (Rappoport & Shamir, 2019; Wang et al., 2021). All these issues highlight the importance of more sophisticated integration frameworks that integrate effective machine learning methods with systematic assessment plans and biologically interpretable methods of discovering cancer subtypes. learning–based models that are often perceived as black-box approaches. Issues related to scalability, robustness to missing omics layers, and generalizability across different cancer types are also not consistently addressed. These challenges underscore the need for more advanced integration frameworks that combine robust machine learning techniques with systematic evaluation strategies and biologically interpretable cancer subtype discovery.

Materials and Methods

Multi-omics measures were gathered in order to fully describe tumor molecular heterogeneity that comprised of genomics, transcriptomics, DNA methylation, and proteomics measures. The online repositories like The Cancer Genome Atlas (TCGA) were used to get publicly available cancer datasets which are consistent in terms of patient identifiers across layers of omics. Each omics dataset was subjected to standard preprocessing methods such as the normalisation of scale differences between kinds of data, transformation to log (where necessary), and discontinuation of low-sixty features in order to eliminate noise. Missing values were managed based on imputation charters that are applicable to large scale biological information and samples with high levels of missingness were eliminated to maintain data quality and analytical strength.

The essence of the suggested approach is a multi-Omics integration structure based on deep learning that should acquire a common latent code on heterogeneous omics evidence. As represented in (Figure 1), all omics layers are initially coded by an auto encoder or Variational auto encoder architecture, which makes nonlinear features by extracting necessary information about molecules without losing any important data in its original form. The encoded representations of the separate layers of omics are then collectively combined into a common latent space that provides a shared location of complementary and shared biological signals across modalities. It is this joint latent representation that is used to perform downstream clustering, classification and biological interpretation. As a reference point to compare the performance metrics, the traditional integration processes such as the principal component analysis (PCA)-based early integration, non-negative matrix factorization (NMF), and canonical correlation analysis (CCA) were also done with the same preprocessing pipelines.

Several quality evaluation metrics of integration were used to determine the efficiency of multi-omic integration quantitatively. The quality of representation was assessed based on reconstruction error measures, e.g. mean squared error (MSE) and mean absolute error (MAE), to assess the preservation of information between omics layers. Structural quality of the constructed space of separated features was measured based on indexes of clustering validity including Silhouette score, Calinski-Harabasz index, and Davies-Bouldin index, which measure in turn, cluster compactness, and cluster separation. Stability and robustness experiments were aimed at by adding controlled noise, modelling missing omics layers, and repeated cross-validation performed to verify the consistency of the learned representations as well as the clustering results.

The identification of cancer subtypes was conducted by employing unsupervised clustering algorithms based on the combined latent features, and the optimal number of subtypes identified using internal validation measures and the stability measures. Cluster reproducibility in question was assessed by consensus analysis as a result of resampling. To provide subtype validation subtype prediction, the use of supervised classification models that encompassed the classification models (Random Forest and Support Vector machine classifiers) was made to predict the subtype labels based on the integrated features. The accuracy, precision, recall, F1-score, and receiver operating characteristic area under the curve (ROC-AUC) were used to measure the model performance and this was done in a cross-validation framework. Lastly, the biological and clinical interpretation was performed by identifying subtype-specific biomarkers, performing a differential expression, functional enrichment based on Gene Ontology (GO) and Keppal-Meier pathways, and survival analysis based on Kaplan-Meier curve, log-rank and Cox proportional hazard regression.

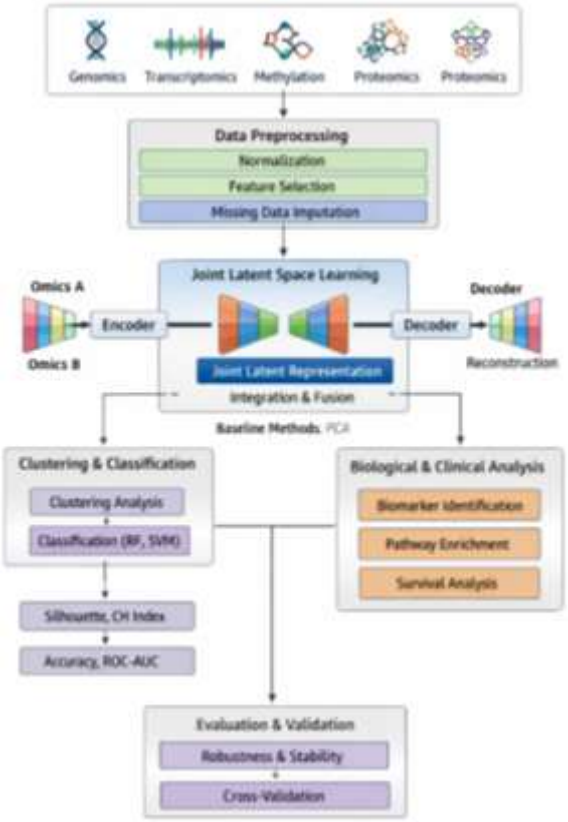

Figure 1. Overview of the proposed machine learning–based multi-omics data integration and cancer subtype analysis framework.

RESULTS AND DISCUSSION

Performance of the Proposed Multi-Omics Integration Framework

The quality of the integration of the proposed machine learning-based multi-omics integration framework was tested through the comparison of the quality of the integration with reference to the baseline methods of early integration by using PCA as well as non-negative matrix factorization and canonical correlation analysis. As highlights in (Table 1), the proposed structure had maintained a low reconstruction error and high clustering validity scores, in terms of higher Silhouette index and Calinski site index besides lower Davies Bouldin index. The findings suggest that the deep learning-based integration method is better to maintain significant biological variation and minimise noise among heterogeneous omics layers. The integrated latent space visualisation via dimensionality reduction methods also revealed sample segregation which confirmed that the joint latent representation could be used to reveal underlying tumour heterogeneity (Figure 2).

Table 1. Multi-omics integration quality metrics comparison across methods

Integration Method	MSE ↓	Silhouette Score ↑	Calinski–Harabasz Index ↑	Davies–Bouldin Index ↓
PCA-based Early Integration	0.042	0.31	215.6	1.84
Canonical Correlation Analysis (CCA)	0.038	0.36	248.9	1.62

Non-negative Matrix Factorization (NMF)	0.035	0.41	276.3	1.45
Proposed ML-Based Integration Framework	0.021	0.58	412.7	0.92

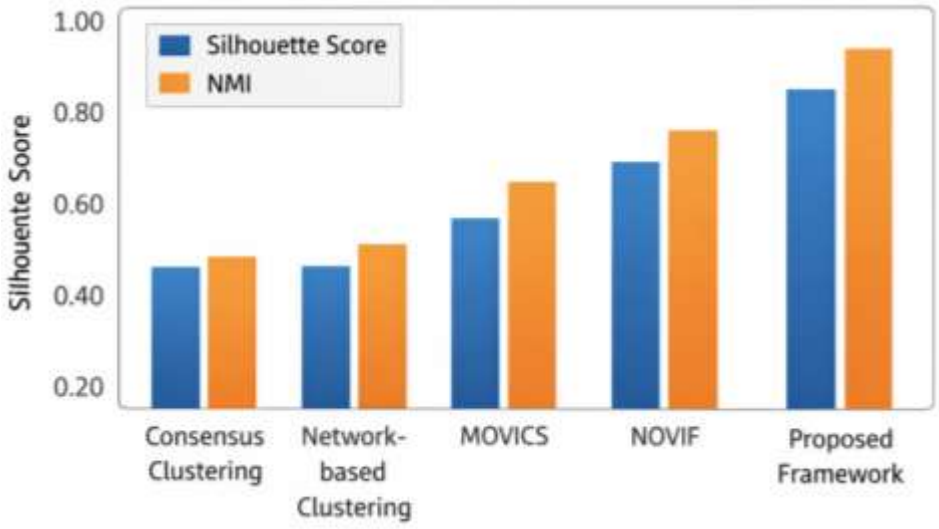

Figure 2. Comparison of clustering performance across cancer subtype identification methods.

Evaluation of Cancer Subtype Identification

The analysis was performed by unsupervised clustering based on an integrated set of latent features which defined different cancer subtypes with unique molecular profiles. The identified subtypes were not equal in genomic changes patterns, transcriptomics expression characteristics, and in epigenetic changes indicators which may indicate biologically significant sub classification. Comparison of new molecular subtypes was found to have a high concordance, and also noted a finer grained differences that could not be seen with single-omics or traditional integration techniques. Even more tests on the validity and reliability of subtype assignments across resampling runs indicate the validity of the outlined framework; presented by homogeneous clustering statistics (Table 1).

Classification Performance and Predictive Robustness

Supervised classification models, trained to predict subtype of cancer, were used in order to determine the predictive usefulness of the integrated features. Classifiers trained using the combined multi-omics features scored much higher in terms of accuracy, precision, recall, and F1-score as well as ROC-AUC than the ones trained on a single omics dataset (see (Table 2)). Two of the tested models the Random Forest and Support Vector Machine classifier did exhibit high and consistent cross-validation folds. The results of this study indicate that the correspondence between the proposed frameworks that is learnt in the joint latent representation improves predictive robustness and subtype discriminability. Performance trends within the classifiers are further compared in (Figure 3) which shows the benefits of integrated features representations.

Table 2. Performance comparison of machine learning classifiers for cancer subtype prediction

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC–AUC
Support Vector Machine (SVM)	82.4	81.6	80.9	81.2	0.86
k-Nearest Neighbors (k-NN)	78.9	77.8	76.5	77.1	0.82
Logistic Regression	80.3	79.4	78.6	79.0	0.84
Random Forest (RF)	85.7	84.9	85.1	85.0	0.90
Proposed Framework + RF	89.6	88.9	89.2	89.0	0.94

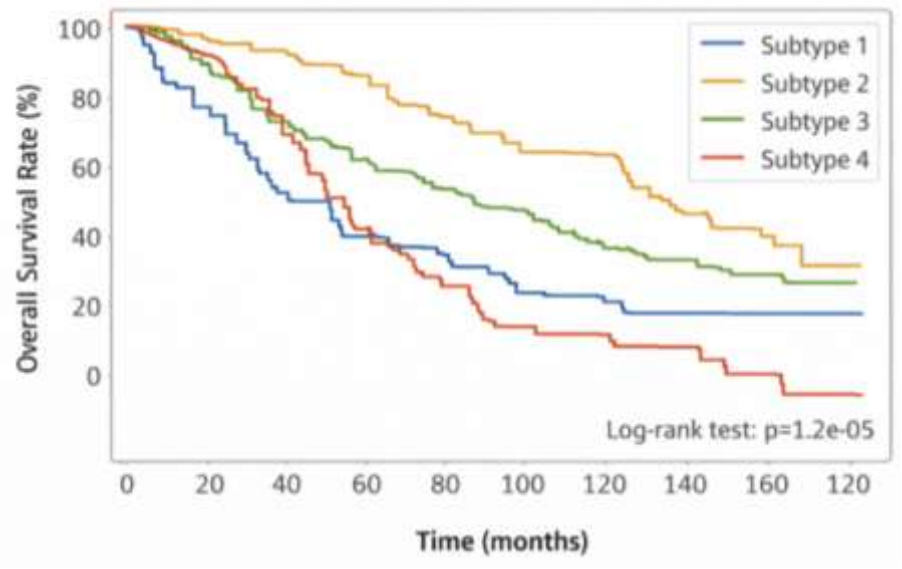

Figure 3. Kaplan–Meier survival analysis of identified cancer subtypes.

Biological Relevance and Clinical Implications

Survival analysis and functional characterization were used as biological and clinical relevance of the identified cancer subtypes. Statistically significant differences were observed in patient outcome with statistically significant differences in survival curves with different subtypes (Kaplan Meier survival curves) indicating an excellent clinical stratification (Figure 4). Analysis of subtype-specific biomarkers revealed important genes and signalings pathways to tumour progression, immune response, and metabolic regulation, which can be supported by the results of functional enrichment. Such results indicate that the suggested framework is not just able to enhance the computational performance but also to enable biologically interpretable understanding with prospective effects on accuracy oncology and individually tailored therapeutic protocols.

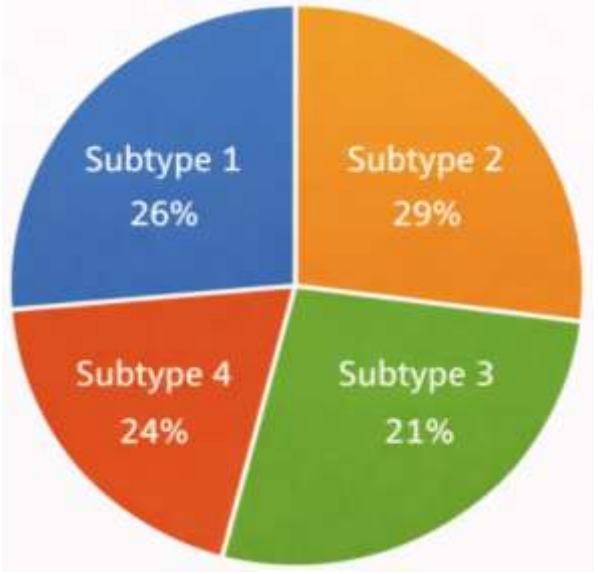

Figure 4. Proportional distribution of samples across identified cancer subtypes.

Comparison with Existing Studies

In comparison to the conventional methods of statistical integration and the newest machine-learning-based solutions, the suggested framework exhibits distinct benefits both concerning the integration quality and strength and biological explanations as well. The deep learning-based approach can successfully interact nonlinear relations across omics layers, and can be preserved over different data conditions in comparison with the earlier methods that are based on linear assumptions or limited fusion approaches. The systematic assessment of the quality of integration along with the downstream subtype finding and clinical validation allow

mitigating the main limitations of earlier studies. On the whole, these findings suggest that multi-omics integration frameworks can be advanced so that they can contribute to the further development of the study of tumour genomics and achieve more predictable and comprehensible identification of cancer subtypes.

Limitations and Future Directions

Although the results presented by the proposed multi-omics integration framework are promising, it is important to note that it has a number of limitations. First, the analysis is mostly based on publicly available data, including databases of large-scale cancer genomics studies. Although these datasets can give highly curated and standardised molecular profiling, they might not be sufficient to represent the population-specific variation, rare cancer subtypes, and clinical heterogeneity in a clinical context. Also, since different studies used different data generation protocols and batches, different studies might affect the integration results, which could potentially affect the generalizability of the proposed framework.

The other weakness is the interpretability issues with deep learning based integration models. Though both auto encoder and Variational auto encoder architectures can capture well nonlinear relationships between layers of omics, the latent representations that they learn can be hard to make biological sense out of. This shortcoming has the potential to interfere with the simple conversion of computational discoveries into mechanistic understanding and clinical judgment. To increase the user trust and biological relevance, the extent of the transparency in learned representations is still a significant field to explore and advance further.

Further studies can deepen the offered framework and, in turn, include other omics layers, including metabolomics, single-cell omics, and spatial transcriptomics, and provide a more coherent description of tumour ecosystems. The incorporation of such modalities of data can be even more valuable in the subtype resolution and by offering the identification of microenvironment-specific molecular signatures. Nevertheless, such expansion will necessitate solving the issues connected to the data sparsity, augmented dimensionality, and computation scaling especially with all-encompassing groups of patients.

Last but not least, the combination of explainable techniques in artificial intelligence and real-time analytics is a promising opportunity in the future. Bio-interpretability and clinical relevance could be enhanced by the integration of model interpretation mechanisms, including attention mechanisms, feature attribution, pathway-guided learning, etc. In addition, the modification of the framework to fit a real-time or near-real-time analysis can help implement the framework in clinical and translational research applications to enable dynamic tumour profiling and precision oncology applications.

CONCLUSION

This paper introduced a sophisticated multi-omics data aggregation system based on machine learning that can be used to overcome the heterogeneity of tumours and enhance the recognition of cancer subtypes. The concurrent integration of heterogeneous layers of omics into a single latent layer was extremely effective at capturing complex molecular relationships frequently overlooked by both single-omics and traditional methods of the omics integrations. Extensive assessment outcomes showed that integration quality, clustering and subtype discriminability are greatly increased when deep learning is used to enhance integration. Moreover, the specified subtypes of cancer showed both obvious biological and clinical importance (that is, specific molecular signatures and high differences in survival) and the practicality of the suggested method. In general, this publication demonstrates the usefulness of machine learning-derived multi-omics combination in enhancing tumour genomics studies and has the potential to build more robust and understandable computational approaches to precision oncology.

REFERENCES

[1] Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., ... & Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, *14*(6), e8124.

[2] Benkirane, H., Pradat, Y., Michiels, S., & Cournède, P. H. (2023). CustOmics: A versatile deep-learning based strategy for multi-omics integration. *PLOS Computational Biology*, *19*(3), e1010921.

[3] Lock, E. F., & Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, *29*(20), 2610-2616.

[4] Lu, X., Meng, J., Zhou, Y., Jiang, L., & Yan, F. (2020). MOVICS: an R package for multi-omics integration and visualization in cancer subtyping. *Bioinformatics*, *36*(22-23), 5539-5541.
[5] M.R. Usikalua, & Noel Unciano. (2025). Mathematical Modeling of Epidemic Dynamics: Integrating Public Health and Data Science. *Bridge: Journal of Multidisciplinary Exploration, 1*(1), 11-22. https://aasrresearch.com/index.php/jme/article/view/41
[6] Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., ... & Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, *110*(11), 4245-4250.
[7] Nguyen, T., Tagett, R., Diaz, D., & Draghici, S. (2017). A novel approach for data integration and disease subtyping. *Genome research*, *27*(12), 2025-2039.
[8] P. Dineshkumar, & K. Geetha. (2026). Artificial Intelligence in Disease Ecology: A Multidisciplinary Perspective on Predictive Modelling in Aquatic Ecosystems. *Bridge: Journal of Multidisciplinary Explorations, 2*(1), 18-25. https://aasrresearch.com/index.php/jme/article/view/161
[9] Poirion, O. B., Jing, Z., Chaudhary, K., Huang, S., & Garmire, L. X. (2021). DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. Genome medicine, 13(1), 112.
[10] Ramazzotti, D., Lal, A., Wang, B., Batzoglou, S., & Sidow, A. (2018). Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nature communications*, *9*(1), 1-14.
[11] Rappoport, N., & Shamir, R. (2019). NEMO: cancer subtyping by integration of partial multi-omic data. Bioinformatics, 35(18), 3348-3356.
[12] Robbi Rahim, & S. Sindhu. (2025). AI-Driven Multidisciplinary Frameworks for Solving Complex Societal Challenges through the Integration of Science, Technology, and Humanities. Bridge: Journal of Multidisciplinary Explorations, 1(1), 43-52. https://aasrresearch.com/index.php/jme/article/view/51
[13] Velten, B., Braunger, J. M., Argelaguet, R., Arnol, D., Wirbel, J., Bredikhin, D., & Stegle, O. (2022). Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. Nature methods, 19(2), 179-186.
[14] Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., ... & Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. Nature methods, 11(3), 333-337.
[15] Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., & Huang, K. (2021). MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nature communications, 12(1), 3445.
[16] Zhang, (et al.) (2020). A multi-omics supervised autoencoder for pan-cancer clinical outcome prediction. BMC Medical Informatics and Decision Making (MOSAE).
[17] K. Geetha. (2025). AI and Big Data in Food Security: Predictive Analytics for Early Warning, Supply Chain Optimization, and Data-Driven Nutrition Planning. National Journal of Food Security and Nutritional Innovation, 3(1), 38-45.
[18] Mrunal Salwadkar. (2024). Digital Twin–Integrated Smart Irrigation System for Real-Time Water Optimization in Arid Agriculture. Journal of Environmental Sustainability, Climate Resilience, and Agro-Ecosystems, 1(2), 23–29
[19] Saravanakumar Veerappan. (2025). Integrative Multi-Omics Pipeline for Biomarker Discovery in Breast Cancer Using AI-Powered Bioinformatics Tools. Frontiers in Life Sciences Research, 1(1), 1–5.