

<Conference abbreviation>

<Conference Series name>

<Volume number and Year> <DOI Number>

Deep Learning-Driven Multi-Omics Integration for Cancer Classification and Biomarker Discovery

Pakhi Sharma¹, Namrata Dhanda², Vandana Yadav³, and Krishna Yadav⁴

^{1,4}Department of CSE, Amity University, Lucknow, U.P., India

²Department of CSE, SAREC, Mirzapur, U.P., India

³Department of CSE, Brainware University, Kolkata, W.B., India

paksh2004@gmail.com | ndhanda510@gmail.com | vandanayadav771@gmail.com | yadav.krishna.tech@gmail.com

Abstract. Breast cancer survival-status classification is difficult because tumour biology varies at several molecular levels at once, which single-omics data rarely captures. This paper integrates copy number variation (CNV), somatic mutation (MU), gene expression (RS) and protein expression (PP) data for survival prediction. Modality-specific autoencoders learn a compact latent representation of each omics layer, and the latent vectors are fused so complementary information is modelled together and used to predict survival status. On a TCGA breast cancer cohort of 705 samples the model reached a test accuracy of 85.1%, an F1-score of 0.222 and a ROC-AUC of 0.582, indicating that latent multi-omics representations carry useful signal while validation on balanced, independent datasets is still needed. Encoder weights also allow candidate biomarkers to be identified.

Keywords. Multi-omics integration; deep learning; cancer classification; autoencoder; biomarker discovery; TCGA; gene expression; copy number variation; proteomics.

1. INTRODUCTION

Breast cancer varies at the genomic, transcriptomic and proteomic levels. High-throughput sequencing has made these changes easier to examine [1], but a single data type gives only a partial view. Combining omics layers allows the disease to be studied from several perspectives and identifies patterns tied to different subtypes and outcomes [2,3]. Because omics datasets carry very large feature counts, machine learning and deep learning are widely applied to cancer classification and survival prediction [2–18]. High dimensionality remains a difficulty, and the models are hard to interpret.

This study integrates CNV, MU, RS and PP data with autoencoders. Because each dataset own structure, a separate autoencoder is trained per omics type, compressing it into a latent representation that retains informative content while suppressing noise [3,4]; the representations are fused and used to predict survival. The contributions are: integration of four omics sources for survival-status classification; modality-specific autoencoders for compact representation learning; a comparison of single-omics, early-fusion and autoencoder-based approaches; and exploration of the learned features for candidate biomarker identification.

2. LITERATURE REVIEW

Azmi et al. [1] compared a stacked denoising autoencoder with a variational autoencoder and assessed feature selection; the variational model reached 91.86% accuracy on lung squamous cell carcinoma. Alharbi et al. [2] applied graph neural networks to biological relationships expressed as graphs, beating conventional methods at higher cost. Subramanian et al. [3] reported that integration generally improves on individual data types, while noting imbalanced data and limited interpretability. Heo et al. [4] found that jointly using genomics, transcriptomics, proteomics and epigenomics improves subtype classification, patient grouping and prognosis, though differences between data types complicate integration. Choi and Chae [5] learned shared patterns across molecular sources for breast cancer, constrained by limited sample numbers.

3. DATASET DESCRIPTION

The data come from The Cancer Genome Atlas (TCGA), widely used for its standardised molecular profiles. Mutation data capture genetic alterations, CNV data show genomic gains and losses, gene expression reflects transcription, and proteomic data describe protein-level changes. The cohort has 705 samples with 1,936 features each (Table 3.1); the target is the “vital status” field, encoded as a binary outcome (Table 3.2).

Tables 3.1, 3.2. Omics distribution (left) and survival status distribution (right).

Omics type	Features	Survival status	Patients
Mutation (MU)	249	Survived (0)	611
Copy Number Variation (CNV)	860	Not survived (1)	94
Gene expression (RS)	604	Total	705
Protein-level profile (PP)	223		

4. METHODOLOGY

4.1. Preprocessing, Scaling and Splitting

The dataset was organised into a separate matrix per omics type, with survival status stored separately to prevent leakage. Z-score normalisation, giving zero mean and unit standard deviation, was applied so no feature dominates training through its range alone:

$$Z = \frac{X - \mu}{\sigma} \quad (4.1)$$

where μ is the feature mean, σ the standard deviation and Z the standardised value. Scaling was applied only to PP and RS: CNV values are discrete codes (0, 1, -1, 2, -2) whose biological meaning scaling would destroy, and MU values are binary. The data were split 70:15:15 into training, validation and test sets by stratified sampling, preserving the class distribution.

4.2. Model Development

Single-omics baselines were trained first with logistic regression, chosen for its interpretability on high-dimensional data. Early fusion followed, concatenating the four matrices into one representation so a model could learn across layers in parallel; logistic regression and random forest were both applied. Such classifiers may miss non-linear relationships, so a deep fusion model was introduced. An autoencoder’s encoder maps input into a smaller latent space and its decoder reconstructs it, so minimising reconstruction error teaches it to keep structure and discard noise. Separate autoencoders were trained on CNV, MU, RS and PP data, their latent features concatenated, and a weight of 1.5 applied to the PP representation before classification (Table 4.1). Features that most influence the prediction, traced through the encoder weights, are treated as candidate biomarkers.

Table 4.1. Autoencoder and training configuration.

Parameter	Configuration	Parameter	Configuration
Encoder	256 neurons, ReLU	Classifier	128 neurons, ReLU
Encoder dropout	0.3	Classifier dropout	0.4
Latent dimension	CNV 64, MU 32, RS 64, PP 32	Output layer	1 neuron, sigmoid
Decoder	256 neurons, ReLU	Optimiser / LR	Adam / 0.001

Parameter	Configuration	Parameter	Configuration
Reconstruction layer	Linear	Batch size / epochs	32 / 50
Fusion dimension	192	Classification loss	Binary cross-entropy
PP latent weight	1.5	Reconstruction loss	MSE (weight 0.05)

5. PERFORMANCE METRICS

Accuracy, F1-score and the area under the receiver operating characteristic curve (ROC-AUC) were used. Accuracy is the proportion of correct predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.1)$$

where TP, TN, FP and FN are true and false positives and negatives. Accuracy misleads under class imbalance, so the F1-score was recorded too; it combines precision (share of predicted positives that are correct) with recall (share of actual positives recovered):

$$Precision = \frac{TP}{TP + FP} \quad (5.2) \quad Recall = \frac{TP}{TP + FN} \quad (5.3)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5.4)$$

ROC-AUC is the area under the curve of true positive rate against false positive rate; higher values indicate stronger, threshold-independent discrimination.

6. RESULTS AND DISCUSSION

Results are summarised in Table 6.1. The dataset is heavily imbalanced, 611 survived against 94 not-survived, so accuracy alone describes performance poorly; the low F1-scores show how hard the minority class is.

Table 6.1. Comparative performance across modelling approaches.

Model	Accuracy	F1-score	ROC-AUC
CNV only (LR)	0.761	0.000	0.349
MU only (LR)	0.717	0.167	0.559
RS only (LR)	0.761	0.261	0.566
PP only (LR)	0.435	0.435	0.716
Early fusion (LR)	0.704	0.087	0.516
Early fusion (RF)	0.887	0.200	0.628
AE fusion (proposed)	0.851	0.222	0.582

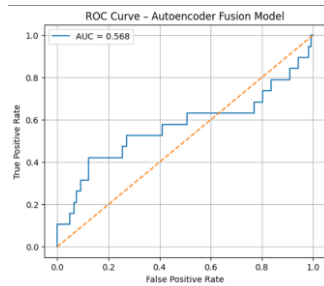


Fig. 6.1. ROC curve for the proposed autoencoder fusion model.

Each omics dataset supplies useful information while fusion gives a broader view, and the latent representations may point to genes and proteins worth following up as biomarkers, subject to independent validation. Limitations: only TCGA data were used; the latent features are hard to interpret; class imbalance affects the F1-score; and the deep model costs more computation.

7. CONCLUSION AND FUTURE WORK

This study built a multi-omics classification framework from copy number variation, somatic mutation, gene and protein expression data in TCGA. The autoencoder fusion model achieved a slightly higher F1-score than the early-fusion random forest, which nonetheless remained ahead on accuracy and ROC-AUC. Future work will test the biomarkers on independent datasets and through pathway analysis, add DNA methylation and miRNA expression, and apply explainable AI to identify which genes and omics layers drive the prediction. Better handling of class imbalance is the most direct route to stronger results.

8. REFERENCES

- [1] N. S. Azmi, et al., ‘Deep learning for cancer classification using multi-omics feature selection’, *Prog. Microbes Mol. Biol.*, vol. 5, no. 1, 2022.
- [2] F. Alharbi, A. Vakanski, B. Zhang, M. K. Elbashir, M. Mohammed, ‘Multi-omics integration using graph neural networks for cancer classification’, *IEEE Access*, vol. 13, pp. 37724–37736, 2025.
- [3] I. Subramanian, et al., ‘Multi-omics data integration, interpretation, and its application’, *Bioinform. Biol. Insights*, vol. 14, 2020.
- [4] Y. J. Heo, et al., ‘Integrative multi-omics in cancer research: networks to clinical subtypes’, *Mol. Cells*, vol. 44, no. 7, pp. 433–443, 2021.
- [5] J. M. Choi, H. Chae, ‘moBRCA-net: breast cancer subtyping using multi-omics attention networks’, *BMC Bioinformatics*, vol. 24, no. 1, 2023.
- [6] Y. Lin, et al., ‘Breast cancer subtypes via deep neural networks on multi-omics data’, *Genes*, vol. 11, 2020.
- [7] Y. Ektefaie, W. Yuan, D. A. Dillon, et al., ‘Integrative multiomics-histopathology analysis for breast cancer classification’, *npj Breast Cancer*, vol. 7, 2021.
- [8] A. F. Abbasi, et al., ‘Multi-omics framework for cancer molecular subtyping’, *Sci. Rep.*, vol. 15, 2025.
- [9] A. M. Hassan, S. M. Naeem, M. A. A. Eldosoky, et al., ‘Multi-omics machine learning for breast cancer subtyping’, *Arab. J. Sci. Eng.*, vol. 50, 2025.
- [10] D. Leng, L. Zheng, Y. Wen, et al., ‘Benchmark of deep learning multi-omics fusion methods for cancer’, *Genome Biol.*, vol. 23, 2022.
- [11] J. Zhang, Y. Che, R. Liu, Z. Wang, W. Liu, ‘Deep learning-driven multi-omics analysis for cancer diagnostics’, *Brief. Bioinform.*, vol. 26, no. 4, 2025.
- [12] D. Wang, J. Gu, ‘Integrative clustering of multi-omics data for cancer classification’, *Quant. Biol.*, vol. 4, 2016.
- [13] J. S. Wekesa, M. Kimwele, ‘Multi-omics deep learning for disease diagnosis’, *Front. Genet.*, vol. 14, 2023.
- [14] B. Ma, et al., ‘Cancer classification using XGBoost and multi-omics data’, *Comput. Biol. Med.*, vol. 121, 2020.
- [15] M. Tao, T. Song, W. Du, S. Han, C. Zuo, Y. Li, Y. Wang, Z. Yang, ‘Breast cancer subtypes using multiple kernel learning on omics data’, *Genes*, vol. 10, 2019.
- [16] L. Tong, J. Mitchel, K. Chatlin, et al., ‘Feature-level integration of multi-omics data for breast cancer survival’, *BMC Med. Inform. Decis. Mak.*, vol. 20, 2020.
- [17] Z. Huang, X. Zhan, S. Xiang, et al., ‘SALMON: survival analysis learning with multi-omics neural networks’, *Front. Genet.*, vol. 10, 2019.
- [18] S. Yang, et al., ‘Machine learning for subtyping triple-negative breast cancer’, *J. Cancer*, vol. 15, 2024.