DeepSSC: A Semi-Supervised Deep Learning Framework for Multi-Omics Cancer Subtype Classification and Biomarker Discovery

Hoang Le[1,†], Van-Minh Nguyen[2,†], Duc-Hai Tran[2], Van-Anh Tran[2], Van-Nui Nguyen[3], Quang-Huy Nguyen[2,4,†,*], Duc-Hau Le[2,4,*]

[1]Faculty of Information Technology, VNU University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam

[2]School of Information and Communiations Technology, Hanoi University of Science and Technology, Hanoi, Vietnam.

[3]Thai Nguyen University of Information and Communication Technology, Thainguyen, 250000, Vietnam

[4]Department of Computational Biomedicine, Vingroup Big Data Institute, Hanoi, Vietnam

† these authors have contributed equally

* corresponding author: Email: huynguyen96.dnu@gmail.com, hauld@soict.hust.edu.vn; Tel: (84)912324564

Keywords: cancer subtyping; breast cancer; lung cancer; colorectal cancer; autoencoder; semi-supervised learning; biomarkers.

ABSTRACT

The rapid advancement of high-throughput -omics technologies has generated large-scale multi-omics datasets, creating new opportunities for integrative disease modeling and biomarker discovery. However, the limited availability of labeled samples relative to model complexity poses challenges for supervised deep learning approaches. To address this issue, we propose DeepSSC, a biphasic deep semi-supervised framework for multi-omics subtype classification and biomarker identification. In Phase 1, modality-specific denoising autoencoders learn compact latent representations for each -omics data type using both labeled and unlabeled samples. These representations are integrated via a post-concatenation strategy and fed into a neural network classifier for subtype prediction. In Phase 2, a Biomarker Gene Identification (BGI) procedure leverages the trained classifier to identify subtype-specific important genes. We evaluated DeepSSC on three cancer types—breast invasive carcinoma (BRCA), colon and rectum adenocarcinoma (COADREAD), and lung cancer—using TCGA datasets and an independent METABRIC validation cohort. DeepSSC consistently outperformed state-of-the-art multi-omics integration methods, particularly when unlabeled samples were incorporated during representation learning. Moreover, the identified biomarkers showed strong concordance with known cancer-associated genes and enriched biologically relevant pathways. Together, these results demonstrate that semi-supervised multi-omics integration improves subtype classification while enabling biologically interpretable biomarker discovery. DeepSSC provides a general and extensible framework for omics-based precision medicine studies.

INTRODUCTION

Cancer is a complex and heterogeneous disease characterized by diverse molecular alterations across multiple biological layers. Advances in high-throughput sequencing technologies have enabled the generation of large-scale multi-omics datasets, including genomics, transcriptomics, epigenomics, proteomics, and other molecular profiles. These complementary data sources provide a comprehensive view of tumor biology and have significantly enhanced our understanding of cancer mechanisms, subtype classification, prognosis prediction, and therapeutic stratification (Andre, et al., 2009; Bhattacharyya, et al., 2015; Blenkiron, et al., 2007; Curtis, et al., 2012; Nguyen, et al., 2020; Shen, et al., 2009).

Despite these advances, effective integration of multi-omics data remains a major computational challenge (Duan, et al., 2021; Nguyen, et al., 2020; Shen, et al., 2012; Subramanian, et al., 2020). Multi-omics datasets are typically high-dimensional, noisy, heterogeneous, and limited in sample size relative to feature space. Differences in scale, distribution, and biological relevance across omics modalities further complicate integrative analysis. Traditional machine learning methods—such as support vector machines, random forests, and logistic regression—often rely on direct feature concatenation or independent modeling strategies (Abbasi, et al., 2025; Acharya and Mukhopadhyay, 2024; Wang, et al., 2023). While these approaches can achieve reasonable performance, they frequently fail to capture complex cross-omics interactions and structured biological relationships.

To address these limitations, deep learning approaches have increasingly been applied to multi-omics integration. Autoencoders and other representation learning frameworks have been widely used to reduce dimensionality and learn latent embeddings from heterogeneous molecular data (Sartori, et al., 2025). More recently, graph neural networks (GNNs) have emerged as powerful tools for modeling structured relationships in biological systems, such as patient similarity networks, gene–gene interaction networks, and cross-omics associations (Corso, et al., 2024; Wu, et al., 2022; Xu, et al., 2019). Recent comprehensive reviews on deep learning and graph neural network–based multi-omics integration for cancer classification highlight that GNN-driven frameworks represent a rapidly growing and promising research direction due to their ability to jointly learn feature representations and relational structures (Alharbi, et al., 2025; Sartori, et al., 2025). Among representative graph-based multi-omics methods, MOGONET (Multi-Omics Graph cOnvolutional NETworks) introduced a supervised framework that constructs omics-specific patient similarity graphs and applies graph convolutional networks to jointly explore omics-specific learning and cross-omics correlation learning (Wang, et al., 2021). MOGONET demonstrated improved classification performance across multiple biomedical tasks and provided a mechanism for identifying important biomarkers from different omics layers. Subsequently, MoGCN proposed a multi-omics integration model that combines autoencoder-based dimensionality reduction and similarity network fusion (SNF) to construct a patient similarity network, which is then processed by a graph convolutional network for cancer subtype classification (Li, et al., 2022). MoGCN achieved strong performance on TCGA breast cancer and pan-kidney datasets and emphasized interpretability through feature extraction and network visualization.

While these approaches represent significant advances in multi-omics integration, several challenges remain. First, many existing frameworks are primarily designed under fully supervised settings, requiring labeled samples for model training (Bu, et al., 2024; Chen, et al., 2024; Choi and Chae, 2023; Li, et al., 2022; Patel, et al., 2025; Wang, et al., 2021). However, in practical biomedical

applications, labeled data are often scarce, whereas unlabeled molecular profiles are abundant (Chen, et al., 2019; Chen, et al., 2023; Chen, et al., 2023; Nguyen, et al., 2020; Yang, et al., 2021)(Cai and Wang, 2024; Pan, et al., 2024; Xie, et al., 2024; Zhao, et al., 2023). Semi-supervised learning strategies have been increasingly recognized as effective approaches for leveraging both labeled and unlabeled samples to enhance representation learning and improve generalization performance, particularly in high-dimensional biomedical contexts (Mastropietro, et al., 2023). Second, multi-omics integration strategies often adopt either early concatenation or fixed fusion mechanisms, implicitly assuming equal contribution from all omics modalities (Duan, et al., 2021). In reality, different omics layers may carry distinct predictive signals depending on cancer type and classification task. Recent reviews emphasize the importance of adaptive or modality-aware integration mechanisms to better balance heterogeneous omics information and avoid dominance or suppression of specific data types (Alharbi, et al., 2025; Sartori, et al., 2025). Third, although biomarker discovery is frequently reported as a downstream analysis, interpretability is not always systematically integrated into model design (Park, et al., 2023). Feature attribution techniques and structured biomarker identification pipelines are increasingly considered essential for translating deep learning models into biologically meaningful insights and clinically actionable knowledge.

Motivated by the challenges of limited labeled samples and heterogeneous multi-omics data integration, we propose DeepSSC, a biphasic semi-supervised framework for cancer subtype classification and biomarker identification. Unlike approaches tailored to a single cancer type or restricted to fully supervised settings, DeepSSC is designed as a generalizable architecture applicable across diverse cancers and data modalities. In this study, we evaluate its performance on three representative cancer types—breast invasive carcinoma (BRCA), colon and rectum adenocarcinoma (COADREAD), and lung cancer—which encompass both multi-class and binary classification settings and diverse molecular characteristics. DeepSSC integrates semi-supervised representation learning with supervised subtype prediction. By adopting a post-concatenation strategy, the framework first learns modality-specific latent representations using denoising autoencoders and then integrates them for classification, enabling effective use of unlabeled samples while preserving omics-specific structure. In addition, a systematic biomarker identification procedure is incorporated to enhance interpretability and biological insight. Through comprehensive evaluation on TCGA datasets and independent validation cohorts, we demonstrate that DeepSSC achieves competitive or superior performance compared with established multi-omics integration methods while providing biologically meaningful subtype-specific biomarkers. This work highlights the value of semi-supervised learning for multi-omics classification and offers a structured, interpretable, and extensible pipeline for precision oncology research.

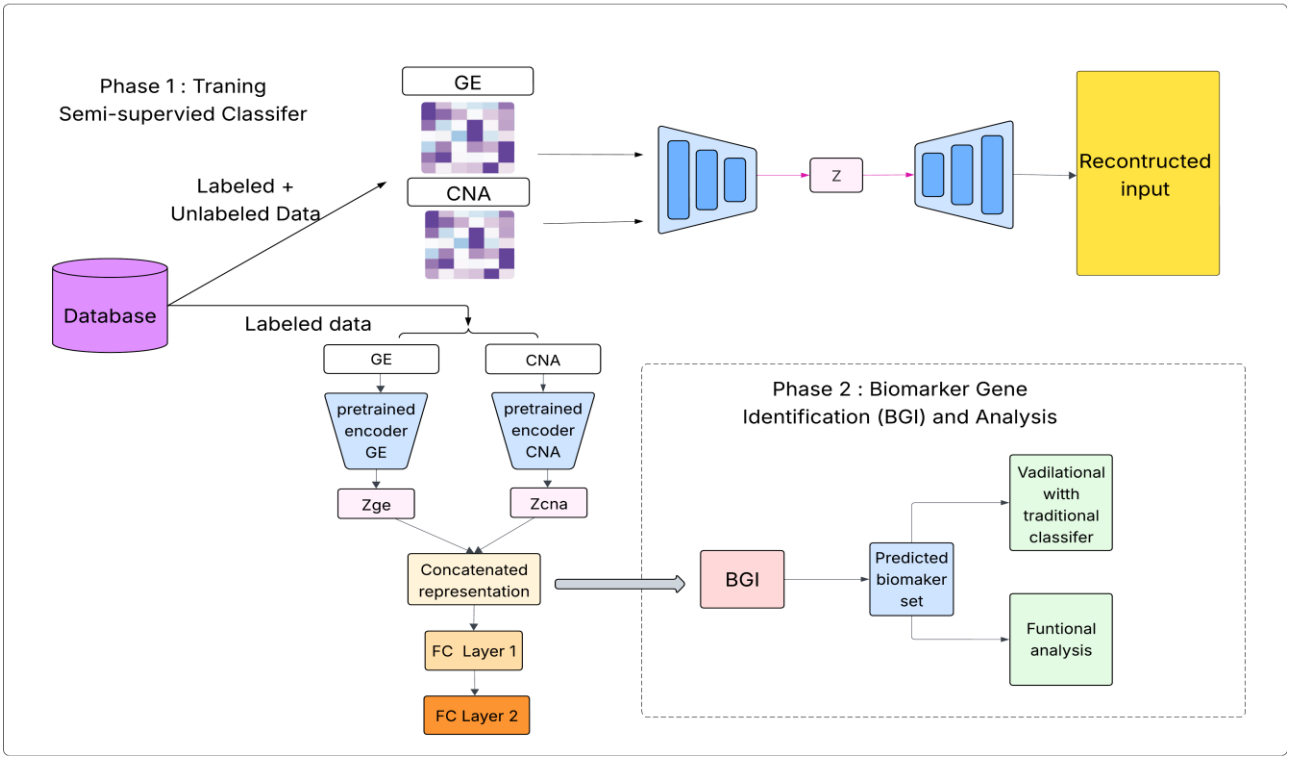

Figure 1 | Overview of the DeepSSC framework. DeepSSC consists of two phases. In Phase 1 (semi-supervised classifier training), preprocessed gene expression (GE) and copy-number alteration (CNA) profiles from TCGA are separately input into two denoising autoencoders (DAEs) to learn omics-specific latent representations, denoted as Z_{ge} and Z_{cna}. These latent features are concatenated and passed through two fully connected (FC) layers to construct a neural network classifier for cancer subtype prediction. In Phase 2 (Biomarker Gene Identification, BGI), the trained classifier is used to identify candidate subtype-specific biomarkers. The resulting biomarker set is evaluated using conventional machine learning models to assess its discriminative power for subtype classification and is further analyzed for biological relevance through functional investigation.

Materials and Methods

Overview of the DeepSSC Framework

In this study, we propose DeepSSC, a biphasic deep semi-supervised multi-omics integration framework designed for two primary tasks: (i) cancer subtype classification and (ii) subtype-specific biomarker identification. DeepSSC is developed as a generalizable framework applicable to multiple biomedical classification problems involving tabular omics data. We evaluate its performance on three representative cancer cohorts, including breast invasive carcinoma (BRCA), colon and rectum adenocarcinoma (COADREAD), and lung cancer. We integrate two complementary omics modalities: mRNA expression (GE) and copy-number alterations (CNA), obtained from The Cancer Genome Atlas (TCGA) (Chang, et al., 2013) and validated using independent datasets from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Pereira, et al., 2016) (Figure 1).

DeepSSC consists of two sequential phases. In Phase 1, two denoising autoencoders (DAEs) are trained independently for GE and CNA profiles to learn omics-specific latent representations. These latent features are subsequently concatenated and passed through two fully connected (FC) layers to construct a semi-supervised neural network classifier for subtype prediction. In Phase 2, we

introduce a Biomarker Gene Identification (BGI) procedure that leverages the trained classifier to identify subtype-specific biomarkers. We demonstrate that the identified biomarkers retain strong discriminative ability when evaluated using conventional machine learning models. Furthermore, we assess their biological relevance through downstream functional analysis. Comparative experiments show that DeepSSC achieves superior or competitive performance relative to state-of-the-art supervised multi-omics classification methods.

Data Collection and Preprocessing

The Cancer Genome Atlas (TCGA) datasets used in this study were downloaded from the UCSC Cancer Genome Browser Xena platform (Goldman, et al., 2020). The METABRIC breast cancer (BRCA) dataset (Pereira, et al., 2016) was obtained from cBioPortal (Cerami, et al., 2012; Gao, et al., 2013) and split into discovery and validation cohorts according to the European Genome-Phenome Archive (EGA; accession number EGAS00000000083). Copy-number alteration (CNA) data measured on the Affymetrix SNP 6.0 platform and gene expression (GE) data generated using the Illumina Human v3 microarray platform were processed to match the formats used for the TCGA datasets.

For TCGA data, invalid samples, including solid tissue normal and metastatic samples, were removed. In addition, male BRCA patients and samples listed in the TCGA blacklist and redaction lists were excluded. Labeled samples were defined as those assigned to a specific disease subtype in both GE and CNA profiles. Unlabeled samples were defined as those assigned to a subtype in only one omics modality or unassigned in both modalities (Figure S1 in Supplementary Materials). Instead of discarding unlabeled samples, we incorporated them into the training of the denoising autoencoders to leverage additional information for representation learning in the TCGA BRCA and TCGA COADREAD datasets.

After preprocessing, the TCGA BRCA dataset contained 819 labeled samples with both omics profiles, along with 259 GE-only and 245 CNA-only unlabeled samples. Among the labeled BRCA samples, 409 were LumA, 187 LumB, 134 Basal-like, 67 HER2-enriched, and 22 Normal-like. For the TCGA COADREAD dataset, we obtained 264 labeled samples with both omics profiles, together with 118 GE-only and 352 CNA-only unlabeled samples. The four consensus molecular subtypes (CMS) were defined according to Guinney et al. (Guinney, et al., 2015). The TCGA lung cancer dataset comprised 1,010 labeled samples with both omics profiles; due to its relatively small sample size, unlabeled samples were not used for this cohort. A summary of subtype distributions for TCGA/METABRIC BRCA, TCGA COADREAD, and TCGA lung cancer is provided in Table 1.
For all TCGA datasets, labeled samples were randomly partitioned into training, validation, and test sets in a 64:16:20 ratio. Data normalization was performed using the mean and standard deviation computed from the training set for the TCGA lung cancer cohort. For the BRCA and COADREAD cohorts, normalization parameters were computed using both labeled training samples and unlabeled samples to support semi-supervised learning. These statistics were subsequently applied to normalize the validation and test sets prior to neural network training.

The same preprocessing pipeline was applied to both the discovery and validation sets of the METABRIC dataset. For genes sharing identical identifiers (i.e., duplicated “Hugo Symbol Entrez” and “Gene_ID” entries), expression values were averaged to obtain a single representative measurement. After harmonization, 17,300 genes were retained for GE data and 22,544 genes for CNA data. For downstream biomarker validation, only labeled METABRIC samples containing biomarkers identified by the proposed BGI procedure (trained on TCGA BRCA) were retained. Three

conventional machine learning classifiers were trained on the discovery cohort, and their predictive performance was evaluated on the independent validation cohort.

	Subtypes	TCGA			METABRIC	
		Train	Validation	Test	Discovery	Validation
BRCA dataset	Basal-like	83	23	28	110	198
	HER2	46	9	12	85	151
	Luminal A	275	61	73	450	246
	Luminal B	110	32	45	261	213
	Normal-like	12	5	5	54	130
	Total	526	130	163	960	938
Lung cancer dataset	ADC	324	86	102	-	-
	SCC	322	76	100	-	-
	Total	646	162	202	-	-
COADREAD dataset	CMS1	24	6	7	-	-
	CMS2	70	18	22	-	-
	CMS3	24	6	8	-	-
	CMS4	50	13	16	-	-
	Total	168	43	53	-	-

Table 1 | Summary of sample distribution across TCGA and METABRIC cohorts, including BRCA, COADREAD, and lung cancer datasets. Abbreviations: BRCA, breast invasive carcinoma; COADREAD, colon and rectum adenocarcinoma; ADC, lung adenocarcinoma; SCC, lung squamous cell carcinoma; TCGA, The Cancer Genome Atlas.

Semi-supervised Neural Network Training and Weighted Loss Function

Denoising Autoencoder Pretraining

Due to the distinct characteristics of gene expression (GE) and copy-number alteration (CNA) data, two independent denoising autoencoders (DAEs) were constructed for each omics modality (Figure 1). Each DAE was trained using both labeled and unlabeled samples to exploit additional information during representation learning.

For GE data, the DAE consisted of five fully connected (FC) layers with exponential linear unit (ELU) activation functions. The ELU activation is defined as:

$$ELU(x) = f(x) = \begin{cases} x, & x > 0 \\ \alpha * (e^x - 1), & x \leq 0 \end{cases} \quad (1)$$

The first three layers formed the encoder, comprising an input layer and two hidden layers with dimensions 20,530 → 4,096 → 2,048. The decoder consisted of two symmetric hidden layers with 4,096 and 20,530 units, respectively.

Similarly, the CNA DAE included five FC layers with dimensions 24,776 → 4,096 → 1,024 → 4,096 → 24,776. Hyperparameters, including the number of hidden units and corruption rate, were tuned to minimize the mean squared error (MSE) on the validation set. The optimal dropout rate applied to the input layer was 0.5.

To mitigate internal covariate shift, batch normalization was applied before the activation function of the first hidden layer in each encoder.

Construction of the Semi-supervised Classifier

The encoder components of the pretrained DAEs were used to initialize the subtype classification network. The latent representations learned from GE (2,048 dimensions) and CNA (1,024 dimensions) were concatenated to form a 3,072-dimensional feature vector. This representation was then passed through an additional hidden layer containing 1,024 neurons.

The final classifier consisted of the two pretrained encoders followed by fully connected layers with ELU activation functions. The last FC layer served as the output layer, producing class scores corresponding to each cancer subtype.

During training, encoder weights were fine-tuned using a smaller learning rate than that applied to the newly added classification layers. This strategy allowed the model to align representations from different omics modalities while preserving learned feature structures. Hyperparameters were selected to maximize validation accuracy, and early stopping was employed to prevent overfitting.

Weighted Cross-Entropy for Class Imbalance

The TCGA BRCA dataset exhibited substantial class imbalance, particularly for the Normal-like subtype, which contained significantly fewer samples than other subtypes. To address this issue, we adopted weighted multi-class cross-entropy loss.

A commonly used balanced heuristic (King and Zeng, 2001) defines the class weight w_i as:

$$w_i = \frac{N}{C \times N_i} \qquad (2)$$

where:

- N is the total number of samples,
- C is the number of classes,
- N_i is the number of samples in class i.

To prevent excessively large weights for minority classes from destabilizing training, we further transformed the weights into a logarithmic scale:

$$w_i = \ln \frac{N}{C \times N_i} \qquad (3)$$

Although both strategies improved performance, we further optimized classification results by treating class weights as tunable hyperparameters. Specifically, weights were manually adjusted based on recall and F1-score performance on the validation set. The weights for Basal-like and HER2 subtypes were fixed at 1, as these classes were comparatively easier to classify. The weights for the remaining subtypes were gradually increased until the average F1-score on the validation set was maximized.

The weighted cross-entropy loss for sample x with ground-truth class i is defined as:

$$\mathcal{L}(x, i) = -w_i log p_i(x) \quad (4)$$

where $p_i(x)$ denotes the predicted probability for class i. The final training loss was computed as the mean loss over each mini-batch.

Biomarker Gene Identification (BGI) Procedure

The Biomarker Gene Identification (BGI) procedure was designed to identify subtype-specific genes that contribute most significantly to the classification decisions of the trained neural network. Specifically, the trained semi-supervised classifier was analyzed using the Captum interpretability library (Kokhlikyan, et al., 2020), a unified framework for interpreting deep learning models.

Among the available attribution methods in Captum, we employed the Integrated Gradients (IG) algorithm to quantify feature importance. For a given target output neuron corresponding to a specific cancer subtype, IG computes the contribution of each input feature (i.e., gene) by integrating gradients along a straight-line path from a baseline input x' to the actual input x. This approach assigns an attribution score to each gene, reflecting its contribution to the model's prediction.

Formally, the integrated gradient for the i i-th feature is defined as:

$$IG_i(x) := (x_i - x_i') \int_0^1 \frac{\partial F\left(x' + \alpha(x - x')\right)}{\partial x_i} d\alpha \quad (5)$$

Where

- $F(\cdot)$ denotes the neural network function,
- x is the input feature vector (GE or CNA profile),
- x' is the baseline input, and
- α parameterizes the straight-line path from x' to x.

In practice, the integral is approximated using a Riemann summation:

$$IG_i(x) \approx (x_i - x_i') \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F\left(x' + \frac{k}{m} \times (x - x')\right)}{\partial x_i} \quad (6)$$

where m denotes the number of approximation steps. This computation is efficiently implemented in modern deep learning frameworks by iteratively evaluating gradients along the interpolation path.

For each subtype, genes were ranked according to their attribution scores, and highly ranked genes were selected as candidate biomarkers.

To evaluate the biological relevance of the identified BRCA biomarkers, we compared them against four established reference sources: the Cancer Gene Census database (Futreal, et al., 2004), the comprehensive genomic analyses reported by Pereira et al. (Pereira, et al., 2016), the study by

Nik-Zainal et al. (Nik-Zainal, et al., 2016), and the PAM50 gene signature (Parker, et al., 2009). A biomarker was considered previously established if it appeared in at least one of these sources. Additional implementation details of the BGI procedure are provided in the Supplementary Methods.

Functional Enrichment Analysis of Predictive Biomarkers

To investigate the biological relevance of the identified predictive biomarkers, functional enrichment analysis was performed using the clusterProfiler R package (Yu, et al., 2012). Gene Ontology (GO) biological process terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were evaluated for overrepresentation among the selected biomarkers.

Statistical significance was assessed using two-sided tests, and resulting *P*-values were adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) correction (Benjamini and Hochberg, 1995). GO terms and KEGG pathways with an adjusted *P*-value ≤ 0.05 were considered significantly enriched.

RESULTS

Semi-supervised Learning Improves Subtype Classification Performance

To evaluate the impact of incorporating unlabeled samples during training, we compared supervised and semi-supervised learning strategies using two sub-datasets derived from the TCGA BRCA cohort. The supervised setting included only labeled samples, whereas the semi-supervised setting leveraged both labeled and unlabeled samples during representation learning, as defined in the Materials and Methods section.

As shown in Figures 2a and 2b, the semi-supervised model consistently outperformed the supervised counterpart across all evaluation metrics. Specifically, integrating unlabeled samples improved overall accuracy from 86.50% to 87.12%, macro-averaged F1-score from 79.76% to 83.43%, and weighted F1-score from 86.30% to 87.10% in the BRCA dataset. Notably, performance gains were more pronounced in minority subtypes, indicating that semi-supervised learning contributed to better representation of underrepresented classes.

A similar trend was observed in the TCGA COADREAD cohort (Figures 2c and 2d). Incorporating unlabeled samples improved overall accuracy from 86.79% to 88.68%, macro F1-score from 85.91% to 85.34% (balanced improvement across classes), and weighted F1-score from 86.81% to 88.82%. These findings demonstrate that leveraging unlabeled molecular profiles enhances the robustness and generalization ability of the model.

Given the substantial class imbalance in the BRCA dataset, particularly for the Normal-like subtype, we evaluated four weighting strategies for the multi-class cross-entropy loss (see Materials and Methods). Among these, the manually tuned weighting scheme achieved the best validation performance. Under this strategy, the integrated GE+CNA model achieved 87.12% accuracy, 83.43% macro F1-score, and 87.10% weighted F1-score across the five PAM50 subtypes (Table 2). Detailed comparisons of the four weighting strategies are provided in Figure S2.

For the COADREAD dataset, we adopted the balanced heuristic weighting strategy, which yielded 88.68% accuracy, 85.34% macro F1-score, and 88.82% weighted F1-score across the four CMS subtypes.

In contrast, the TCGA lung cancer cohort exhibited minimal class imbalance; therefore, no weighting strategy was applied. The integrated GE+CNA model achieved an accuracy of 96.04%, an F1-score of 95.96%, and an AUC of 0.9924, indicating strong discriminative performance (Table 1).

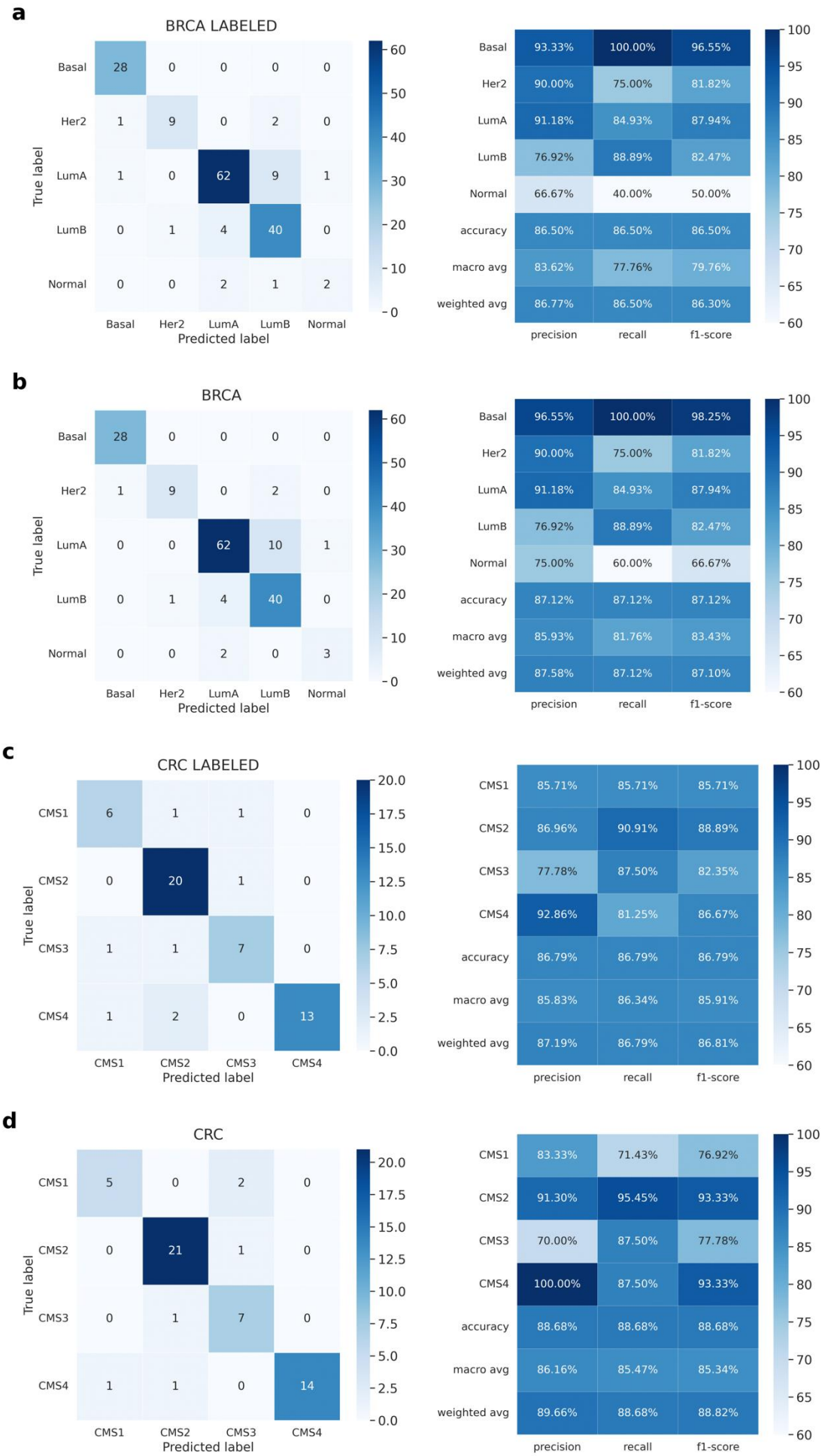
a
BRCA LABELED
Basal 28 0 0 0 0
Her2 1 9 0 2 0
LumA 1 0 62 9 1
LumB 0 1 4 40 0
Normal 0 0 2 1 2
True label
Basal Her2 LumA LumB Normal
Predicted label
60 50 40 30 20 10 0
Basal 93.33% 100.00% 96.55%
Her2 90.00% 75.00% 81.82%
LumA 91.18% 84.93% 87.94%
LumB 76.92% 88.89% 82.47%
Normal 66.67% 40.00% 50.00%
accuracy 86.50% 86.50% 86.50%
macro avg 83.62% 77.76% 79.76%
weighted avg 86.77% 86.50% 86.30%
precision recall f1-score
100 95 90 85 80 75 70 65 60
b
BRCA
Basal 28 0 0 0 0
Her2 1 9 0 2 0
LumA 0 0 62 10 1
LumB 0 1 4 40 0
Normal 0 0 2 0 3
True label
Basal Her2 LumA LumB Normal
Predicted label
60 50 40 30 20 10 0
Basal 96.55% 100.00% 98.25%
Her2 90.00% 75.00% 81.82%
LumA 91.18% 84.93% 87.94%
LumB 76.92% 88.89% 82.47%
Normal 75.00% 60.00% 66.67%
accuracy 87.12% 87.12% 87.12%
macro avg 85.93% 81.76% 83.43%
weighted avg 87.58% 87.12% 87.10%
precision recall f1-score
100 95 90 85 80 75 70 65 60
c
CRC LABELED
CMS1 6 1 1 0
CMS2 0 20 1 0
CMS3 1 1 7 0
CMS4 1 2 0 13
True label
CMS1 CMS2 CMS3 CMS4
Predicted label
20.0 17.5 15.0 12.5 10.0 7.5 5.0 2.5 0.0
CMS1 85.71% 85.71% 85.71%
CMS2 86.96% 90.91% 88.89%
CMS3 77.78% 87.50% 82.35%
CMS4 92.86% 81.25% 86.67%
accuracy 86.79% 86.79% 86.79%
macro avg 85.83% 86.34% 85.91%
weighted avg 87.19% 86.79% 86.81%
precision recall f1-score
100 95 90 85 80 75 70 65 60
d
CRC
CMS1 5 0 2 0
CMS2 0 21 1 0
CMS3 0 1 7 0
CMS4 1 1 0 14
True label
CMS1 CMS2 CMS3 CMS4
Predicted label
20.0 17.5 15.0 12.5 10.0 7.5 5.0 2.5 0.0
CMS1 83.33% 71.43% 76.92%
CMS2 91.30% 95.45% 93.33%
CMS3 70.00% 87.50% 77.78%
CMS4 100.00% 87.50% 93.33%
accuracy 88.68% 88.68% 88.68%
macro avg 86.16% 85.47% 85.34%
weighted avg 89.66% 88.68% 88.82%
precision recall f1-score
100 95 90 85 80 75 70 65 60

Figure 2 | Semi-supervised learning improves subtype classification performance in BRCA and COADREAD cohorts. Confusion matrices (left) and corresponding classification metrics (right) for models trained under supervised and semi-supervised settings. (**a**) BRCA using only labeled samples (supervised learning). (**b**) BRCA using both labeled and unlabeled samples (semi-supervised learning). (**c**) COADREAD using only labeled samples. (**d**) COADREAD using both labeled and unlabeled samples. For each cancer type, the model architecture and hyperparameter settings were identical across experimental conditions. Performance was evaluated using accuracy, precision, recall, and F1-score.

Impact of Multi-omics Integration on Classification Performance

To assess the contribution of multi-omics integration, we compared the classification performance of DeepSSC using three input configurations: (i) integrated GE+CNA data, (ii) GE-only data, and (iii) CNA-only data. For the single-omics settings, the model architecture was modified by removing the encoder corresponding to the excluded modality, followed by hyperparameter re-tuning on the validation set before evaluation on the test set. As shown in Table 2, integrating GE and CNA consistently improved overall classification performance compared with single-omics models across most datasets and evaluation metrics.

For the TCGA BRCA cohort, the integrated model achieved the best performance across all metrics, with an accuracy of 87.12%, macro F1-score of 83.43%, and weighted F1-score of 87.10%. In contrast, the GE-only model achieved slightly lower performance (86.50% accuracy), while the CNA-only model exhibited substantially reduced predictive ability (61.96% accuracy), indicating that CNA alone was insufficient for robust subtype discrimination in BRCA.

In the TCGA COADREAD dataset, multi-omics integration improved accuracy from 86.79% (GE-only) to 88.68% and increased weighted F1-score from 86.83% to 88.82%. Although the macro F1-score for GE-only (85.98%) was marginally higher than that of the integrated model (85.34%), the overall performance gain in accuracy and weighted F1-score suggests that incorporating CNA data enhanced classification stability across subtypes. The CNA-only model performed poorly (37.74% accuracy), further emphasizing the complementary nature of GE and CNA information.

For the TCGA lung cancer cohort, the integrated model achieved 96.04% accuracy and 95.96% F1-score, outperforming both GE-only (93.56% accuracy) and CNA-only (89.11% accuracy) configurations. The AUC values were high across all settings (>97%), indicating strong discriminative ability; however, multi-omics integration yielded the best overall classification metrics.

Importantly, across all three cancer types, the GE-only model consistently outperformed the CNA-only model, highlighting the dominant predictive contribution of gene expression data. Nevertheless, integrating CNA with GE further enhanced performance in most cases, demonstrating that CNA provides complementary subtype-specific information that improves overall predictive robustness.

Collectively, these results confirm that multi-omics integration within DeepSSC offers measurable advantages over single-omics approaches, particularly in terms of accuracy and weighted F1-score, while maintaining strong performance across diverse cancer types.

	TCGA BRCA			TCGA lung cancer			TCGA COADREAD		
	ACC	macro F1	weighted F1	ACC	F1	AUC	ACC	Macro F1	Weighted F1
GE+CNA	**87.12%**	**83.43%**	**87.10%**	**96.04%**	**95.96%**	99.24%	**88.68%**	85.34%	**88.82%**
GE	86.50%	82.83%	86.28%	93.56%	93.33%	**99.26%**	86.79%	**85.98%**	86.83%
CNA	61.96%	41.66%	60.36%	89.11%	88.54%	97.04%	37.74%	34.72%	35.85%

Table 2 | Comparison of classification performance using different omics modalities across TCGA BRCA, lung cancer, and COADREAD datasets. Performance of DeepSSC when trained with integrated multi-omics data (GE+CNA) versus single-omics inputs (GE only or CNA only). Evaluation metrics include accuracy (ACC), macro-averaged F1-score (Macro F1), weighted F1-score (Weighted F1), and area under the ROC curve (AUC, reported for lung cancer). Abbreviations: GE, gene expression; CNA, copy-number alterations; ACC, accuracy; AUC, area under the receiver operating characteristic curve.

Comparison with State-of-the-Art Methods

To evaluate whether the architectural design of DeepSSC is optimal, we compared it with three pre-concatenation variants and two state-of-the-art multi-omics integration methods. Specifically, we implemented the following variations: (i) pre-concat-AE, where multi-omics features were concatenated before being input into a standard autoencoder (AE); (ii) pre-concat-DAE, where concatenated features were fed into a denoising autoencoder (DAE); and (iii) pre-concat-VAE, where concatenated features were input into a variational autoencoder (VAE). All variants followed the same semi-supervised training strategy and used identical network depths and hidden dimensions as the DAE component in DeepSSC to ensure a fair comparison.

We further compared DeepSSC with two recently proposed supervised multi-omics integration frameworks: MOGONET (Wang, et al., 2021) and MoGCN (Li, et al., 2022). MOGONET employs graph convolutional networks for omics-specific representation learning and a View Correlation Discovery Network to model cross-omics interactions at the label space. It also uses an ablation-based feature masking strategy for biomarker identification, which typically requires multi-stage feature preselection. MoGCN similarly integrates multi-omics data using graph convolutional networks.

Performance was evaluated using accuracy (ACC), weighted F1-score, and macro F1-score for the TCGA BRCA and COADREAD datasets, and ACC, F1-score, and AUC for the TCGA lung cancer dataset. In Phase 2 of DeepSSC, only the top-10 important genes per subtype (46 genes for BRCA, 38 for COADREAD, and 14 for lung cancer) were used to train classical classifiers (SVM, LR, and RF).

Comparison with Pre-concatenation Variants

As shown in Table 3, DeepSSC consistently outperformed its three pre-concatenation variants across datasets. For example, on the TCGA BRCA dataset, DeepSSC (Phase 1 with unlabeled samples) achieved 87.1% accuracy and 83.4% macro F1-score, whereas pre-concat-AE, pre-concat-DAE, and pre-concat-VAE obtained lower macro F1-scores (77.7%, 78.3%, and 74.6%, respectively). Similar trends were observed for COADREAD, where DeepSSC substantially exceeded all pre-concatenation models.

These results demonstrate that the post-concatenation strategy adopted in DeepSSC—learning modality-specific latent representations before integration—is more effective than directly concatenating raw multi-omics features under the semi-supervised learning setting.

Comparison with MOGONET and MoGCN

DeepSSC also showed competitive or superior performance compared with MOGONET and MoGCN across the three cancer types.

For TCGA BRCA, DeepSSC (Phase 1 with unlabeled samples) achieved 87.1% accuracy, outperforming MOGONET (82.2%) and slightly exceeding MoGCN (86.5%). Notably, DeepSSC obtained substantially higher macro F1-score (83.4%) compared with MoGCN (76.8%) and MOGONET (72.0%), indicating improved balanced performance across subtypes.

For TCGA lung cancer, DeepSSC achieved 96.0% accuracy and F1-score, exceeding both MOGONET (94.6%) and MoGCN (94.6%). The AUC of DeepSSC reached 99.2%, comparable to MoGCN (99.1%) and higher than MOGONET (97.5%), confirming strong discriminative capability.

The most pronounced improvement was observed in the TCGA COADREAD dataset. DeepSSC (Phase 1 with unlabeled samples) achieved 88.6% accuracy and 88.8% weighted F1-score, whereas MoGCN obtained 69.9% accuracy and 66.8% weighted F1-score, and MOGONET achieved 84.9% accuracy. This substantial margin highlights the robustness of the semi-supervised post-concatenation integration strategy, particularly in datasets with limited sample sizes.

In Phase 2, DeepSSC-derived biomarkers were used as inputs to conventional machine learning classifiers (SVM, LR, RF). Remarkably, even with a small number of selected genes, these classifiers achieved performance comparable to or exceeding MOGONET and MoGCN across datasets. For example, on COADREAD, RF achieved 92.4% accuracy, surpassing all deep learning baselines.

This finding indicates that DeepSSC not only improves predictive performance through semi-supervised multi-omics integration but also identifies highly discriminative and compact biomarker sets that generalize well to classical classifiers. Overall Conclusion Collectively, the results demonstrate that: (i) The post-concatenation semi-supervised architecture of DeepSSC is superior to pre-concatenation alternatives. (2) DeepSSC achieves competitive or superior performance compared with graph-based state-of-the-art methods. (3) The biomarkers identified by DeepSSC retain strong predictive power even when used with simple machine learning models. These findings validate both the architectural design and the semi-supervised integration strategy of DeepSSC for robust multi-omics cancer subtype classification.

			TCGA BRCA			TCGA Lung Cancer			TCGA COADREAD		
			ACC	weighted F1	macro F1	ACC	F1	AUC	ACC	Weighted F1	Macro F1
DeepSSC	Phase 1	Without unlabeled samples	86.5	86.3	79.8	**96.0**	**96.0**	**99.2**	86.8	86.8	85.9
		With unlabeled samples	**87.1**	**87.1**	**83.4**	-	-	-	**88.6**	**88.8**	**85.3**
	Phase 2	SVM	**91.4**	**91.6**	**89.7**	95.5	95.4	**98.9**	88.6	88.4	86.1
		LR	88.9	89.0	88.6	**96.0**	**96.0**	98.8	88.6	88.7	88.6
		RF	87.7	87.7	84.0	95.0	94.9	**98.9**	**92.4**	**92.2**	**92.2**
pre-concat-AE			85.3	85.0	77.7	-	-	-	79.3	79.3	78.9
pre-concat-DAE			85.2	85.0	78.3	-	-	-	79.3	79.3	78.9
pre-concat-VAE			81.6	81.1	74.6	-	-	-	77.4	77.0	75.1
MOGONET			82.2	81.3	72.0	94.6	94.5	97.5	84.9	84.5	82.7
MOGCN			86.5	85.6	76.8	94.6	94.6	99.1	69.9	66.8	59.8

Table 3 | Performance comparison of DeepSSC and state-of-the-art multi-omics integration methods for TCGA BRCA subtype classification, TCGA COADREAD subtype classification, and TCGA lung cancer type classification. DeepSSC results are reported for Phase 1 (semi-supervised training with and without unlabeled samples) and Phase 2, where only the top-10 important genes per subtype were used to train classical classifiers (SVM, LR, and RF). The best performance for each evaluation metric within each dataset is highlighted in bold.

Identified Biomarkers Retain Predictive Power with Classical Classifiers

To investigate whether the biomarkers identified by DeepSSC retain predictive power independent of the deep learning framework, we evaluated their classification performance using traditional machine learning models. For each cancer type, we selected the top-100 important genes per subtype, resulting in 428 unique genes for BRCA, 359 for COADREAD, and 127 for lung cancer (allowing overlap among subtypes). The complete ranked gene lists are provided in Table S1.

We then assessed subtype discrimination performance using three classical classifiers: Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF). To further evaluate modality contribution, we compared models trained on integrated GE+CNA data, GE-only data, and CNA-only data using LR. Performance was evaluated using repeated 5-fold cross-validation (10 repetitions), and the model with the best overall performance across folds was tested on held-out data.

The three cancer types represent both multi-class classification problems (BRCA with five PAM50 subtypes; COADREAD with four CMS subtypes) and binary classification (lung cancer with two histologic subtypes), allowing comprehensive evaluation.

BRCA Subtype Classification (TCGA and METABRIC)

We first evaluated whether the BRCA subtype-specific biomarkers identified by DeepSSC retained predictive power when used with classical machine learning models. Using the top-ranked genes (top-5 to top-100 per PAM50 subtype; 428 unique genes in total), we trained Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) classifiers.

Performance on TCGA BRCA As shown in Figure 3a and Table S2, all three classifiers achieved strong subtype discrimination across gene set sizes. Even small gene subsets yielded high accuracy

(>81%), indicating that the highest-ranked biomarkers already contained substantial discriminative information. Performance generally improved as more top genes were included, plateauing thereafter. The best accuracy reached approximately 92% across classifiers (SVM with top-10 genes, RF with top-30 genes, and LR with top-60 genes).

These results demonstrate that DeepSSC successfully prioritizes highly informative subtype-specific biomarkers that remain effective even when used in simple linear or tree-based models.

To assess the contribution of different omics modalities, we compared LR models trained using integrated GE+CNA data, GE-only data, and CNA-only data (Figure 3b and Table S3). Models using GE-only data performed comparably to those using integrated data across most thresholds, while CNA-only models showed markedly reduced accuracy and macro F1 scores. This suggests that gene expression provides the dominant subtype-discriminative signal in BRCA, whereas CNA contributes complementary but weaker independent predictive information.

External Validation on METABRIC BRCA To evaluate generalizability, we applied the same ranked gene sets to the independent METABRIC cohort. As shown in Figure 3c and Table S4, classification performance remained robust despite differences in patient populations and experimental platforms. The best LR model achieved approximately 75% accuracy, and all classifiers exhibited performance improvements as the number of top genes increased. These findings indicate that the identified biomarkers are not cohort-specific and retain predictive value across datasets.

Figure 3d and Table S5 further examine modality effects in METABRIC using LR. Consistent with TCGA results, GE-only and integrated GE+CNA models showed similar performance, while CNA-only models performed substantially worse. The similarity of performance trends between TCGA and METABRIC underscores the stability and reproducibility of the identified biomarker sets.

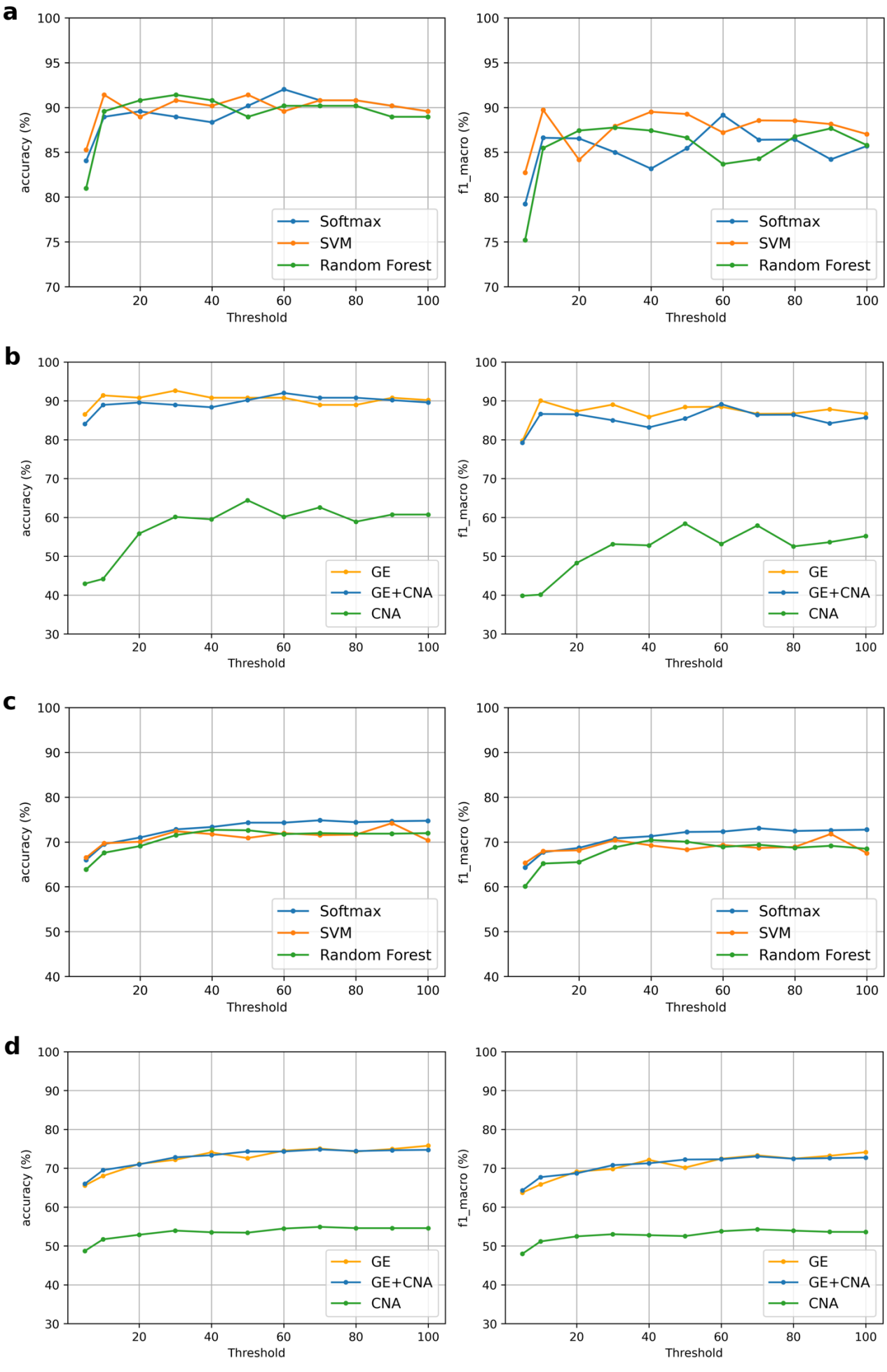
a
accuracy (%)
f1_macro (%)
Threshold
Softmax
SVM
Random Forest
b
GE
GE+CNA
CNA
c
d

Figure 3 | Predictive performance of DeepSSC-identified biomarkers for BRCA subtype classification in TCGA and METABRIC cohorts. (**a**, **c**) Classification performance of three classical classifiers (Logistic Regression, Support Vector Machine, and Random Forest) using top-ranked subtype-specific genes (top-5 to top-100 per PAM50 subtype) as input features on TCGA BRCA (a) and METABRIC BRCA (c). (**b**, **d**) Performance comparison of Logistic Regression models trained using integrated GE+CNA data, GE-only data, and CNA-only data on TCGA BRCA (b) and METABRIC BRCA (d). Models were trained using repeated 5-fold cross-validation (10 repetitions). The model achieving the highest macro-averaged F1-score across cross-validation folds was evaluated on the TCGA test set and the independent METABRIC validation set.

COADREAD Subtype Classification

We next evaluated the predictive capability of identified biomarkers on the TCGA COADREAD dataset (Figure 4), another multi-class classification task. Across all top-gene thresholds, classification accuracy exceeded 86% for all three classifiers (Figure 4a and Table S6). The LR classifier achieved approximately 98% accuracy using the top-50 genes, comparable to SVM and RF, demonstrating that even simple linear models can effectively discriminate CMS subtypes when using DeepSSC-derived biomarkers. Modality comparison (Figure 4b and Table S7) revealed that GE-only and integrated GE+CNA models performed similarly and substantially better than CNA-only models, consistent with observations in BRCA. These results indicate that DeepSSC identifies highly informative gene sets that are robust across multi-class classification settings.

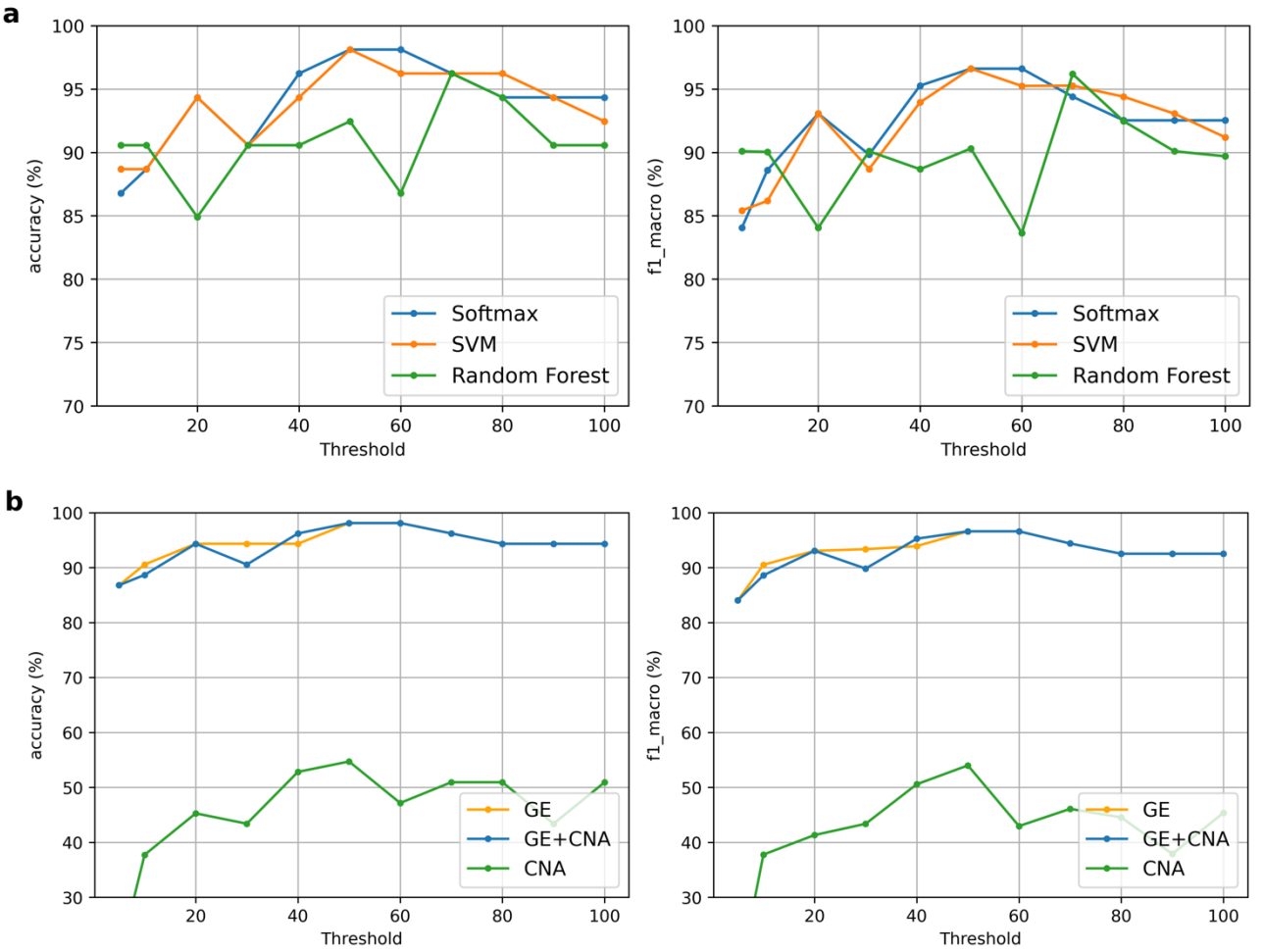

Figure 4 | Predictive performance of DeepSSC-identified biomarkers for COADREAD subtype classification. (**a**) Classification performance of Logistic Regression, Support Vector Machine, and

Random Forest using top-ranked CMS subtype-specific genes (top-5 to top-100 per subtype) as input features. (**b**) Performance comparison of Logistic Regression models trained using integrated GE+CNA data, GE-only data, and CNA-only data across different top-gene thresholds. Models were trained using repeated 5-fold cross-validation (10 repetitions), and the model with the highest overall accuracy across folds was evaluated on the held-out test set.

Lung Cancer Type Classification

Finally, we evaluated lung cancer subtype discrimination (adenocarcinoma vs. squamous cell carcinoma), representing a binary classification task (Figure 5). Remarkably, using only the top-five genes per subtype (six unique genes in total), all three classifiers achieved excellent performance, exceeding 94.5% accuracy and 98.4% AUC (Figure 5a and Table S8). Increasing the number of top genes yielded only marginal improvements, with performance fluctuations generally within 1–2% for accuracy and <1% for AUC. This indicates that DeepSSC successfully identifies a highly compact yet powerful biomarker set for lung cancer classification. Modality analysis (Figure 5b and Table S9) again showed that GE-only and integrated GE+CNA models performed comparably and substantially better than CNA-only models. These findings reinforce the dominant predictive contribution of gene expression data while confirming the robustness of the identified biomarkers across data modalities.

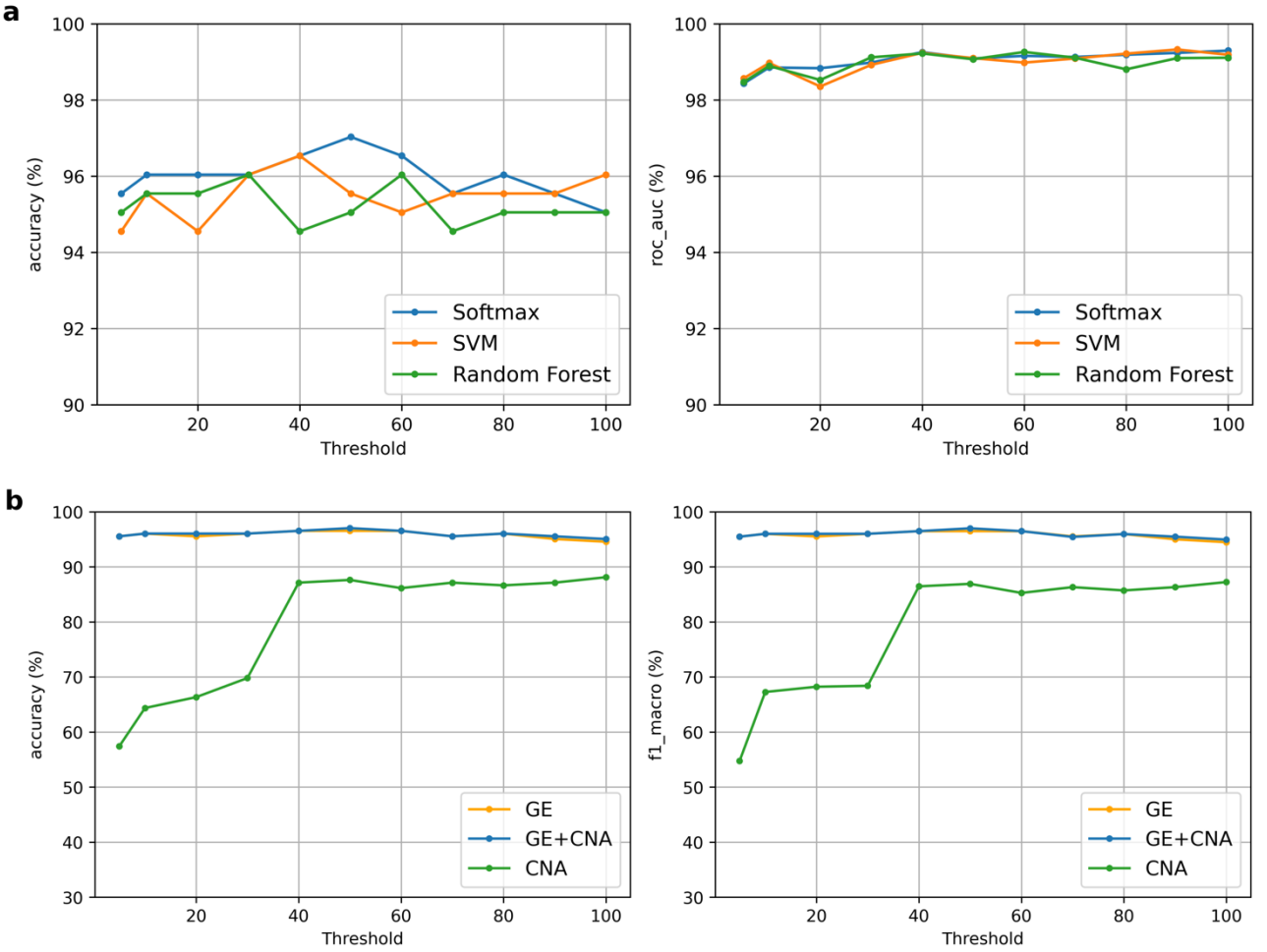

Figure 5 | Predictive performance of DeepSSC-identified biomarkers for lung cancer subtype classification. (**a**) Classification performance of Logistic Regression, Support Vector Machine, and Random Forest using top-ranked histologic subtype-specific genes (top-5 to top-100 per subtype) as input features. (**b**) Performance comparison of Logistic Regression models trained using integrated GE+CNA data, GE-only data, and CNA-only data across different top-gene thresholds. Models were trained using repeated 5-fold cross-validation (10 repetitions), and the model with the highest overall

accuracy across folds was evaluated on the held-out test set. Accuracy and AUC are reported for this binary classification task.

Biological Validation of Identified Biomarkers

To further evaluate the biological relevance of the biomarkers identified by DeepSSC, we compared the top-ranked subtype-specific genes with well-established BRCA-associated gene resources and performed functional enrichment analyses as described in the *Materials and Methods* section.

Concordance with Established BRCA-Associated Genes

Table 4 summarizes the overlap between the top-100 genes in each BRCA subtype and four well-known sources of BRCA-associated genes, including the Cancer Gene Census (Chaudhary, et al., 2018), Nik-Zainal et al. (Nik-Zainal, et al., 2016), Pereira et al. (Pereira, et al., 2016), and the PAM50 gene signature (Parker, et al., 2009). Several well-established breast cancer genes were highly ranked by DeepSSC, including GRB7, ERBB2, and PIK3R1, demonstrating strong agreement with previously reported driver and subtype-defining genes.

To maintain consistency with the PAM50 framework proposed by Parker et al. (Parker, et al., 2009), we selected the top-10 genes per intrinsic subtype (bold red genes in Table S1a), resulting in 46 unique genomic biomarkers, with limited overlap among subtypes. Figure 6a presents the heatmap of expression levels of these 46 biomarkers across the five BRCA molecular subtypes. Three genes—ERBB2, MIA, and GRB7—are included in the established PAM50 signature (Parker, et al., 2009), further supporting the biological validity of our ranked gene list.

Notably, ERBB2 and GRB7, computationally assigned to the HER2 subtype by DeepSSC, are well known to be co-amplified and co-overexpressed in HER2-positive breast tumors (Chen, et al., 2016; Kauraniemi, et al., 2001). Their high ranking and subtype specificity in our model reflect well-established molecular characteristics of HER2-enriched tumors. These observations indicate that DeepSSC effectively recovers canonical subtype-defining genes without prior biological supervision. Functional enrichment analysis of the 46 BRCA biomarkers revealed significant associations with cancer-related pathways (Table S10), including the p53 signaling pathway (hsa04115; adjusted P-value = 0.03) and prostate cancer pathway (hsa05215; adjusted P-value = 0.04). The enrichment of core oncogenic pathways further confirms that the identified biomarkers are not only predictive but also biologically meaningful in the context of tumorigenesis.

COADREAD Biomarkers and CMS-Associated Biological Processes

Using the same selection strategy for COADREAD, we extracted the top-10 genes per CMS subtype (bold red genes in Table S1b), resulting in 38 unique genomic biomarkers with limited overlap across subtypes. The expression heatmap in Figure 6b illustrates clear subtype-specific expression patterns across the four CMS groups.

Functional enrichment analysis (Table S11) identified significant associations with immune- and inflammation-related pathways, including type I diabetes mellitus (hsa04940; adjusted P-value = 0.04) (González, et al., 2017), autoimmune thyroid disease (hsa05320; adjusted P-value = 0.04) (Duran, et al., 2012; L'Heureux, et al., 2019), and inflammatory bowel disease (hsa05321; adjusted P-value = 0.04) (Stidham and Higgins, 2018). These findings are consistent with the known immunological and inflammatory characteristics of colorectal cancer, particularly the immune-enriched CMS1 subtype. Therefore, DeepSSC-derived biomarkers capture biologically coherent processes that reflect established subtype-specific mechanisms.

Lung Cancer Biomarkers and Histologic Differentiation

For lung cancer, we selected the top-10 genes per histologic subtype (bold red genes in Table S1c), resulting in 14 unique genomic biomarkers. The heatmap shown in Figure 6c demonstrates clear separation between lung adenocarcinoma (ADC) and squamous cell carcinoma (SCC) based on the expression patterns of these genes.

Enrichment analysis (Table S12) identified keratinocyte differentiation (GO:0030216; adjusted P-value = 0.008) (Park, et al., 2017), a biological process closely associated with squamous epithelial biology, as well as cytolysis (GO:0019835; adjusted P-value = 0.01), which plays an important role in immune responses and has been implicated in lung cancer treatment strategies (Degos, et al., 2019). These results indicate that the identified biomarkers reflect key biological differences between ADC and SCC and are consistent with known histologic and immunological distinctions.

Across BRCA, COADREAD, and lung cancer, the biomarkers identified by DeepSSC demonstrate strong biological plausibility. They recover well-established subtype-defining genes reported in prior studies (Chaudhary, et al., 2018; Nik-Zainal, et al., 2016; Parker, et al., 2009; Pereira, et al., 2016), exhibit clear subtype-specific expression patterns, and enrich for biologically meaningful cancer-related pathways. Collectively, these findings indicate that DeepSSC not only achieves strong predictive performance but also identifies biologically interpretable and mechanistically relevant biomarkers across diverse cancer types.

Top-100 genes in each subtype	**Cancer Gene Census**	**Nik Zainal *et al.***	**Pereira *et al*.**	**PAM50 genes**
Basal	FOXA1, GATA3	FOXA1, GATA3	GATA3	FOXA1, MAPT, MIA
Her2	ERBB2	ERBB2, FGFR2	ERBB2	BCL2, ERBB2, GRB7
LumA	-	-	-	ANLN, BCL2, EXO1, MAPT, MELK
LumB	-	-	-	ANLN, CDC6, KRT17, KRT5, MIA
Normal	-	PIK3R1	-	-

Table 4 | Overlap between DeepSSC-identified BRCA subtype-specific biomarkers and established BRCA-associated gene resources. Comparison of the top-100 genes identified for each BRCA intrinsic subtype with genes reported in the Cancer Gene Census (Chaudhary, et al., 2018), Nik-Zainal et al. (Nik-Zainal, et al., 2016), Pereira et al. (Pereira, et al., 2016), and the PAM50 gene signature (Parker, et al., 2009). The presence of well-established subtype-defining genes among the top-ranked biomarkers demonstrates concordance between DeepSSC predictions and previously validated BRCA-associated genes.

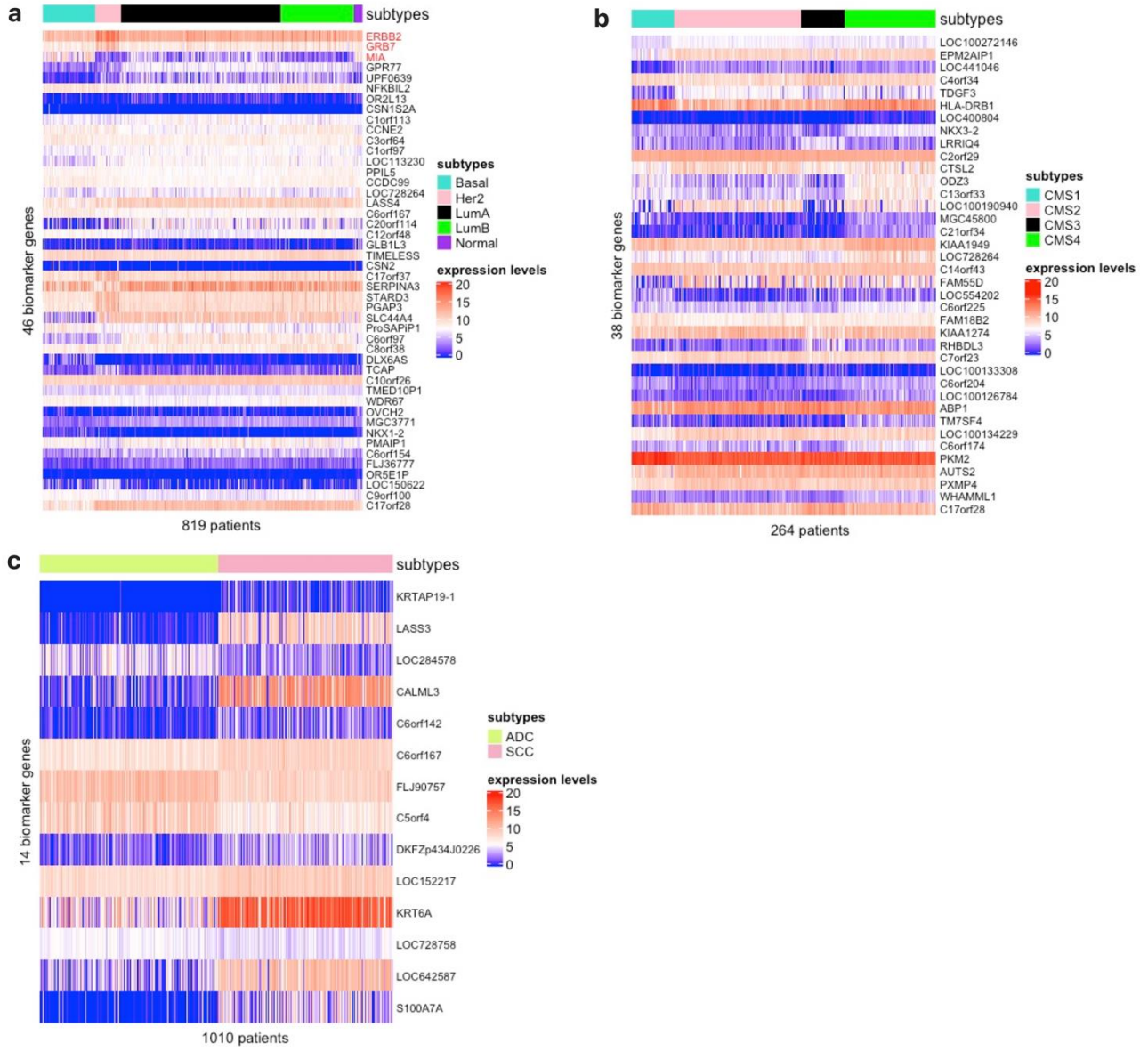

Figure 6 | Expression patterns of DeepSSC-identified subtype-specific biomarkers across cancer types. (**a**) Heatmap showing expression levels of 46 BRCA biomarkers (top-10 genes per PAM50 subtype) across the five intrinsic subtypes in TCGA BRCA. Gene names highlighted in red indicate inclusion in the PAM50 signature (Parker, et al., 2009). (**b**) Heatmap of 38 COADREAD biomarkers (top-10 genes per CMS subtype) across the four consensus molecular subtypes (CMS1–CMS4). (**c**) Heatmap of 14 lung cancer biomarkers (top-10 genes per histologic subtype) across lung adenocarcinoma (ADC) and lung squamous cell carcinoma (SCC) samples. Color scales represent normalized gene expression levels. These heatmaps illustrate clear subtype-specific expression patterns identified by DeepSSC.

Conclusions and Discussion

The high cost of data collection and annotation, together with incomplete knowledge of disease subtyping, has historically limited the availability of labeled biomedical datasets. As a result, many early multi-omics integration studies primarily relied on unsupervised learning strategies, clustering samples without incorporating phenotypic labels and subsequently interpreting the biological relevance of the derived groups. However, with the rapid advancement of high-throughput -omics technologies and the establishment of large-scale cancer genomic resources such as TCGA and METABRIC, multi-omics datasets with detailed clinical annotations have become increasingly accessible. These developments enable supervised and semi-supervised approaches that directly leverage subtype labels to improve predictive modeling and biomarker discovery.

Despite this progress, the number of labeled samples remains relatively small compared to the learning capacity of modern deep learning architectures. To address this challenge, we proposed DeepSSC, a biphasic deep semi-supervised multi-omics integration framework for subtype classification and biomarker identification. DeepSSC treats each -omics modality as a distinct aspect of the same biological entity and performs omics-specific representation learning using denoising autoencoders (DAEs). The learned latent representations are then integrated through a multi-layer perceptron for subtype prediction. This post-concatenation strategy enables effective multi-aspect learning while maintaining interpretability through downstream biomarker importance analysis.

Our results demonstrate that incorporating unlabeled samples during representation learning significantly improves classification performance. The comparison between DeepSSC trained with and without unlabeled data confirms the advantage of semi-supervised learning for multi-omics integration. Furthermore, comparisons with alternative integration strategies show that the post-concatenation approach outperforms pre-concatenation variants. A plausible explanation is that pre-concatenation may introduce bias toward dominant modalities and may fail to fully exploit unlabeled samples when certain -omics types are missing. In contrast, learning modality-specific latent representations before integration allows each data type to contribute effectively while preserving samples that contain only partial multi-omics profiles.

Beyond predictive performance, DeepSSC provides biologically meaningful subtype-specific biomarkers via the BGI procedure. The identified biomarkers not only enable accurate subtype discrimination using conventional machine learning classifiers but also show strong concordance with known cancer-associated genes and pathways. In BRCA, several well-established genes such as ERBB2, GRB7, FOXA1, and GATA3 were highly ranked, and enrichment analyses revealed cancer-related pathways including the p53 signaling pathway. Similar biological consistency was observed for COADREAD and lung cancer, where the selected biomarkers were enriched in disease-relevant pathways and functional processes. These findings support both the robustness and interpretability of the proposed framework.

Nevertheless, several limitations warrant discussion. As observed in the BRCA experiments, DeepSSC—like many existing predictors—has difficulty clearly separating Luminal A and Luminal B tumors. One potential factor is class imbalance in the dataset. Moreover, prior studies have suggested that the distinction between LumA and LumB may not represent entirely coherent biological groups (Weigelt, et al., 2010). The inconsistency across predictors further supports this notion. Previous efforts, including our earlier work (Nguyen, et al., 2020) and that of Netanely et al. (Netanely, et al., 2016), have attempted to recluster luminal tumors to refine subtype boundaries. Historically, luminal BRCA subtypes have undergone multiple redefinitions, indicating that this remains an open biological and computational question. Future work should therefore explore

improved subtype refinement strategies, potentially integrating additional molecular layers or adopting adaptive labeling frameworks.

Another limitation is that the present study focused on gene expression (GE) and copy number alteration (CNA) data in three cancer types (BRCA, COADREAD, and lung cancer). Although the framework is general and can be extended to additional -omics modalities—such as DNA methylation, mutation profiles, or proteomics—and to other cancers (e.g., kidney cancer, brain lower grade glioma) or complex diseases (e.g., Alzheimer's disease), further validation across diverse datasets will strengthen its generalizability.

In conclusion, DeepSSC provides an effective semi-supervised deep learning framework for multi-omics subtype classification and biomarker discovery. By integrating modality-specific representation learning with supervised classification, it achieves strong predictive performance while retaining biological interpretability. Our findings highlight the importance of leveraging unlabeled data in -omics-based predictive modeling and suggest that semi-supervised multi-omics integration should be more broadly considered in future precision medicine research.

DATA AVAILABILITY AND CODE REPORTING

The raw data from TCGA and METABRIC used in the study are available in the UCSC Cancer Genome Browser Xena (https://xenabrowser.net/) and the cBioportal website: (http://www.cbioportal.org), respectively. Approval by a local ethics committee was not required, and all the data can be immediately downloaded to serve for research purposes. We also made them available by pushing to this (https://www.kaggle.com/dataset/a63f5d7411c02e9aa222ddc53384d342a721b06c2d59cba94e1bb68cf66a535f). Python and R codes for reproducing all results are hosted on Github (https://github.com/hauldhut/DeepSSC).

REFERENCES

Abbasi, A.F., *et al.* Multi-omics driven computational framework for cancer molecular subtype classification. *Scientific Reports* 2025;15(1):44141.

Acharya, D. and Mukhopadhyay, A. A comprehensive review of machine learning techniques for multi-omics data integration: challenges and applications in precision oncology. *Briefings in Functional Genomics* 2024;23(5):549-560.

Alharbi, F., *et al.* Comparative Analysis of Multi-Omics Integration Using Graph Neural Networks for Cancer Classification. *IEEE Access* 2025;13:37724-37736.

Andre, F., *et al.* Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2009;15(2):441-451.

Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995;57(1):289-300.

Bhattacharyya, M., Nath, J. and Bandyopadhyay, S. MicroRNA signatures highlight new breast cancer subtypes. *Gene* 2015;556(2):192-198.

Blenkiron, C., *et al.* MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome biology* 2007;8(10):R214.

Bu, Y., *et al.* Cancer molecular subtyping using limited multi-omics data with missingness. *PLOS Computational Biology* 2024;20(12):e1012710.

Cai, Y. and Wang, S. Deeply integrating latent consistent representations in high-noise multi-omics data for cancer subtyping. *Briefings in Bioinformatics* 2024;25(2).

Cerami, E., *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery* 2012;2(5):401-404.

Chang, K., *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* 2013;45(10):1113-1120.

Chaudhary, K., *et al.* Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clinical Cancer Research* 2018;24(6):1248-1259.

Chen, F., *et al.* Supervised graph contrastive learning for cancer subtype identification through multi-omics data integration. *Health Information Science and Systems* 2024;12(1):12.

Chen, R., *et al.* Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. *Bioinformatics* 2019;36(5):1476-1483.

Chen, W., Wang, H. and Liang, C. Deep multi-view contrastive learning for cancer subtype identification. *Briefings in Bioinformatics* 2023;24(5).

Chen, Y., *et al.* Gene expression inference with deep learning. *Bioinformatics* 2016;32(12):1832-1839.

Chen, Y., *et al.* MOCSS: Multi-omics data clustering and cancer subtyping via shared and specific representation learning. *iScience* 2023;26(8).

Choi, J.M. and Chae, H. moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks. *BMC Bioinformatics* 2023;24(1):169.

Corso, G., *et al.* Graph neural networks. *Nature Reviews Methods Primers* 2024;4(1):17.

Curtis, C., *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 2012;486(7403):346-352.

Degos, C., *et al.* Endometrial Tumor Microenvironment Alters Human NK Cell Recruitment, and Resident NK Cell Phenotype and Function. *Frontiers in immunology* 2019;10:877-877.

Duan, R., *et al.* Evaluation and comparison of multi-omics data integration methods for cancer subtyping. *PLOS Computational Biology* 2021;17(8):e1009224.

Duran, C., *et al.* Frequency of Thyroid Nodules among Patients with Colonic Polyps. *Gastroenterology Research and Practice* 2012;2012:178570.

Futreal, P.A., *et al.* A census of human cancer genes. *Nat Rev Cancer* 2004;4(3):177-183.

Gao, J., *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling* 2013;6(269):pl1.

Goldman, M.J., *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnology* 2020;38(6):675-678.
González, N., *et al.* 2017 update on the relationship between diabetes and colorectal cancer: epidemiology, potential molecular mechanisms and therapeutic implications. *Oncotarget* 2017;8(11):18456-18485.
Guinney, J., *et al.* The consensus molecular subtypes of colorectal cancer. *Nature Medicine* 2015;21(11):1350-1356.
Kauraniemi, P., *et al.* New amplified and highly expressed genes discovered in the ERBB2 amplicon in breast cancer by cDNA microarrays. *Cancer Res* 2001;61(22):8235-8240.
King, G. and Zeng, L. Logistic Regression in Rare Events Data. *Political Analysis* 2001;9(2):137-163.
Kokhlikyan, N., *et al.* Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896* 2020.
L'Heureux, A., *et al.* Association Between Thyroid Disorders and Colorectal Cancer Risk in Adult Patients in Taiwan. *JAMA network open* 2019;2(5):e193755-e193755.
Li, X., *et al.* MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis. *Frontiers in Genetics* 2022;Volume 13 - 2022.
Mastropietro, A., De Carlo, G. and Anagnostopoulos, A. XGDAG: explainable gene–disease associations via graph neural networks. *Bioinformatics* 2023;39(8).
Netanely, D., *et al.* Expression and methylation patterns partition luminal-A breast tumors into distinct prognostic subgroups. *Breast Cancer Research* 2016;18(1):74.
Nguyen, Q.-H., *et al.* Multi-omics analysis detects novel prognostic subgroups of breast cancer. *Frontiers in Genetics* 2020.
Nik-Zainal, S., *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 2016;534(7605):47-54.
Nik-Zainal, S., *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 2016;534(7605):47-54.
Pan, L., *et al.* DEDUCE: Multi-head attention decoupled contrastive learning to discover cancer subtypes based on multi-omics data. *Computer Methods and Programs in Biomedicine* 2024;257:108478.
Park, H.J., *et al.* Keratinization of lung squamous cell carcinoma is associated with poor clinical outcome. *Tuberculosis and respiratory diseases* 2017;80(2):179-186.
Park, J., Lee, J.W. and Park, M. Comparison of cancer subtype identification methods combined with feature selection methods in omics data analysis. *BioData Mining* 2023;16(1):18.
Parker, J.S., *et al.* Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology* 2009;27(8):1160-1167.
Patel, J.C., *et al.* GAIN-BRCA: a graph-based AI-net framework for breast cancer subtype classification using multiomics data. *Bioinformatics Advances* 2025;5(1).
Pereira, B., *et al.* The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature Communications* 2016;7(1):11479.
Sartori, F., *et al.* A Comprehensive Review of Deep Learning Applications with Multi-Omics Data in Cancer Research. *Genes* 2025;16(6):648.
Shen, R., *et al.* Integrative Subtype Discovery in Glioblastoma Using iCluster. *PLOS ONE* 2012;7(4):e35236.
Shen, R., Olshen, A.B. and Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics (Oxford, England)* 2009;25(22):2906-2912.
Stidham, R.W. and Higgins, P.D.R. Colorectal Cancer in Inflammatory Bowel Disease. *Clin Colon Rectal Surg* 2018;31(3):168-178.
Subramanian, I., *et al.* Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and biology insights* 2020;14:1177932219899051-1177932219899051.
Wang, S., Wang, S. and Wang, Z. A survey on multi-omics-based cancer diagnosis using machine learning with the potential application in gastrointestinal cancer. *Frontiers in Medicine* 2023;Volume 9 - 2022.

Wang, T., *et al.* MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications* 2021;12(1):3445.

Weigelt, B., *et al.* Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol* 2010;11(4):339-349.

Wu, L., *et al.* Graph neural networks: foundation, frontiers and applications. In, *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2022. p. 4840-4841.

Xie, M., *et al.* Subtype-MGTP: a cancer subtype identification framework based on multi-omics translation. *Bioinformatics* 2024;40(6).

Xu, K., *et al.* How Powerful are Graph Neural Networks? In, *International Conference on Learning Representations*. 2019.

Yang, H., *et al.* Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data. *Bioinformatics* 2021;37(16):2231-2237.

Yu, G., *et al.* clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology* 2012;16(5):284-287.

Zhao, J., *et al.* Subtype-DCC: decoupled contrastive clustering method for cancer subtype identification based on multi-omics data. *Briefings in Bioinformatics* 2023;24(2).