

Original Paper

Deep Generative and Graph-Based Representation Learning for Multiomics Survival Stratification in Ovarian Cancer: Secondary Analysis

Carlos Marino^{1,2}, PhD; Claudia Diaz Paz^{1,2}, DBA, MPhil

¹Pontificia Universidad Católica del Perú, Lima, Peru

²CENTRUM Católica Graduate Business School, Lima, Peru

Corresponding Author:

Carlos Marino, PhD
CENTRUM Católica Graduate Business School
Jirón Daniel Alomía Robles 125
Lima 15023
Peru
Phone: 51 626 7100
Email: cmarino@pucp.pe

Abstract

Background: Ovarian cancer remains one of the most lethal gynecologic malignancies, largely due to pronounced molecular heterogeneity, nonspecific clinical presentation, and frequent diagnosis at advanced stages. Multiomics profiling—including genomics, transcriptomics, and epigenomics—offers a powerful avenue for characterizing this complexity and enabling more precise patient stratification.

Objective: This study aimed to address key challenges in multiomics analysis, including high dimensionality, cross-modality heterogeneity, limited sample size, and the lack of effective approaches for survival stratification of patients with ovarian cancer through deep representation learning.

Methods: We analyzed multiomics data from The Cancer Genome Atlas and developed a 5-stage deep learning pipeline centered on variational autoencoders (VAEs) for nonlinear dimensionality reduction and latent representation learning. A graph convolutional neural network component is described as a proposed extension for modeling interaction-aware representations but was not empirically evaluated in this study. Latent embeddings derived from the VAE were clustered using k-means, and their prognostic relevance was assessed using Cox proportional hazards modeling and Kaplan-Meier survival analysis.

Results: Following correction of a clinical-molecular harmonization issue, the final matched cohort comprised 291 patients. Silhouette analysis identified k=2 as the optimal clustering solution (silhouette=0.272). Kaplan-Meier analysis demonstrated significantly different overall survival between the two clusters (log-rank $\chi^2_1=10.0$; $P=.002$). Cox proportional hazards modeling estimated a hazard ratio of 0.519 (95% CI 0.343-0.785; $P=.002$), indicating that patients assigned to cluster 1 exhibited an approximately 48% lower hazard of death than those in cluster 0. These results demonstrate that the learned latent representations capture prognostically relevant structure within the integrated multiomics data.

Conclusions: The proposed VAE-based framework identified 2 prognostically distinct patient subgroups with significantly different overall survival. These findings demonstrate the potential of deep generative representation learning for multiomics-based survival stratification in ovarian cancer and provide a foundation for future validation in independent cohorts and the evaluation of graph-based extensions.

JMIR Bioinform Biotech 2026;7:e89069; doi: [10.2196/89069](https://doi.org/10.2196/89069)

Keywords: ovarian cancer; personalized medicine; multiomics; variational autoencoders; graph convolutional neural networks; survival analysis

Introduction

Background

Ovarian cancer ranks eighth among cancers affecting women worldwide in both incidence and mortality [1,2], with an estimated 324,603 new cases and 206,956 deaths in 2022 [1, 2]. Although other cancers affecting women, including breast and cervical cancer, have a higher global incidence, ovarian cancer remains a major cause of cancer-related mortality [1, 2]. Ovarian cancer is often asymptomatic in early stages [3,4] and presents nonspecific symptoms in subsequent stages [3]. Thus, early-stage detection, although desirable, is difficult to achieve.

Cancer involves several alterations and interactions, including genetic, epigenetic, and metabolic [5]. Thus, omics and multiomics technologies are a promising source for early-stage diagnosis of cancer, including ovarian cancer, [4]; they could guide prevention, risk-reducing surgeries, and treatments [4]. In addition, genomic data and their use have been growing [6,7]. “Omics” comes from the Latin word *omnis*, which means “everything” [8], and “describes a comprehensive quantitative characterization of a class of molecules in a given biological sample or specimen, aiming to understand the molecular mechanisms and underpinnings underlying the functioning of an organism” [9].

Omics-based studies and the identification of biomarkers provide an important foundation for the development of personalized medicine approaches in cancer and can promote the development of personalized medicine [10,11]. Personalized medicine differs from randomized controlled trials [10]. The latter solve the problem of bias with randomization but provide results based on averages and confidence levels, whereas personalized medicine aims to deal with individual patients’ characteristics [10]. Personalized medicine uses biological information and biomarkers on a molecular level to provide tailored preventive and therapeutic solutions [12]. Moreover, this tailoring approach can also include additional clinical information to improve these solutions [13]. It encompasses not only diagnosis and treatment but also substages such as prognosis, presymptomatic testing, and risk and recurrence assessment [7].

The literature highlights the advantages of a multiomics approach over a mono-omics one in cancer research [14, 15]. A multiomics design provides a more comprehensive perspective and insights on the interactions of different types of omics [14]. Specifically, “multiomics approaches in clinical oncology integrate data from various molecular levels to enhance precision medicine” [14]. On this basis, studies on different types of cancer, such as gastric cancer [16,17], pancreatic cancer [18], breast cancer [19-22], pancancer [23-25], and ovarian cancer, have adopted a multiomics perspective.

In this context, the role of AI and machine learning (ML) has been highlighted as they can assess the large volumes of data that a multiomics study entails [14]. A multiomics ML approach provides a better understanding

of the phenomena but, at the same time, introduces several challenges. Multiomics studies require integrating imbalanced [26], heterogeneous [4], and high-dimensional datasets [3, 27] with limited sample sizes [3,13,27] that could lead to overfitting, limited generalizability, and inaccurate results [15] and capture interactions among genes [28]. In addition, different ML techniques have been useful to predict survival groups in ovarian cancer, with those based on deep learning being the most promissory ones [15,29,30].

On the basis of this, the focus of the current study was on assessing ovarian cancer multiomics data that differentiate survival [3,15] risk groups [31]. To do so, an ensemble of ML techniques was used. Graph neural networks (GNNs) and graph convolutional neural networks (GCNNs) allow for the capture of interactions among different omics [28,32], their biological associations, and patterns of cooperation among features [33]. Data augmentation has been proposed as one strategy to address limited sample sizes in deep learning [13]. In the present study, however, variational autoencoders (VAEs) were used exclusively for nonlinear dimensionality reduction and latent representation learning.

Literature Review

Predicting survival rates of cancer is an important topic that could allow for the better management of this disease [34] and healing therapies [34]. These predictions could include overall survival, recurrence survival, progression survival, or treatment response [30]. In addition, different types of ML algorithms have been used for this purpose, including models such as random forest, support vector machines, Extreme Gradient Boosting (XGBoost), and deep learning techniques such as GNNs [30] and GCNNs. In addition, the literature acknowledges the Cox proportional hazards (CoxPH) model and its variations as an important and widely used framework [34].

Several studies have aimed to make prognoses or predict survival rates or survival groups in different types of cancer, including ovarian cancer. Among those studies that adopted a multiomics approach, such as the current inquiry, some of them assessed the problem of data [15,20,29,35]. Zhang et al [29] used principal component transformation for dimensionality treatment to improve the performance of the model. In their study, deep learning algorithms obtained better predictive power than decision trees and random forest. Cox regression was used to identify molecular features related to patients’ survival rates.

There are several techniques to deal with the so-called curse of dimensionality in cancer-related studies. Some of them are component or factor based [36], such as principal component analysis [37-39] and principal component transformation [29]. Other methods are projection based, such as isometric mapping, uniform manifold approximation and projection, or t-distributed stochastic neighbor embedding [36]. Modern complex techniques such as autoencoders and VAEs have also been applied for data integration [40] and dimensionality reduction [35,41-43]. The literature points out that there are several variations of VAEs [6]. Similar to autoencoders, a VAE encompasses an encoder and a decoder.

The encoder, through multiple layers, produces an embedding vector in the latent space; this latent vector is then decoded, seeking to reconstruct the input data as well as possible [6], reducing their dimensionality. In addition, the latent space of a VAE is regulated based on a known distribution [6].

Studies that predicted survival rates of ovarian cancer using multiomics data have also used autoencoders or VAEs for data integration and dimensionality reduction [15,35]. In the study by Jiang et al [15], a multilayer perceptron was subsequently used to construct a prognostic prediction index for stratifying patients with ovarian and breast cancer into high- and low-risk groups. This model relied on deep learning methods and not on traditional ones such as CoxPH and its variations, random survival forest, XGBoost, and XGBoost with accelerated failure time. Hira et al [35] also propounded a special type of VAE—maximum mean discrepancy VAE—to integrate multiomics data while treating their imbalance, heterogeneity, and high dimensionality. The compressed features of the model were used to cluster samples into molecular subtypes and cancer or noncancer groups. Then, an artificial neural network classified cancer samples and molecular subtypes. The survival analysis included identifying survival subgroups based on univariate CoxPH and clustering samples using the k-means clustering algorithm. The model then predicted survival groups with a support vector machine-based classifier and made a potential prognosis of biomarkers. Jiang et al [15] and Hira et al [35] used the dataset from The Cancer Genome Atlas (TCGA).

The literature has remarked on the relevance of GNNs for biological data assessment. GNNs can model the complexity of data [44], including molecular data such as omics, biological processes, and their interactions [45]. GNNs combine graph structures with the prediction power of deep learning techniques and have different derivations, such as GCNNs [46]. Graph structures are a combination of nodes and edges, where edges provide information about the relationships between nodes [46]. Node classification, edge

classification, and link prediction are three main functions that have been related to GNNs [44]. Survival analysis has been one of the uses of GNNs in the topic of cancer [34,46].

In the field of ovarian cancer survival prediction, GNNs have been used with omics data [32] and other types of information, such as laboratory information, vital signs, and treatment records [34]. Wang et al [32] propounded a prognosis model, DFASGCNS (Dual Fusion Channels and Stacked Graph Convolutional Neural Network), for ovarian cancer based on a stacked GCNN and dual-fusion channels. The stacked GCNN allowed the model to better depict the interactions of multiomics data. This study also used the TCGA dataset.

Studies about cancer survival prediction have propounded ensemble models that combine GNNs or GCNNs and VAEs to achieve more robust results [20,47]. Breast cancer was the topic of one of these inquiries [20]. This study used three ensemble models: long short-term memory, VAE, and GCNN. They were optimized using stochastic gradient descent. The VAE compressed the high-dimensional data into a lower-dimensional space. The dataset was taken from the Molecular Taxonomy of Breast Cancer International Consortium. The model obtained 98% accuracy for the optimized long short-term memory model. In addition, Zhang et al [47] sought to improve the accuracy of GNNs while facing the problem of the limited number of neighboring genes in the data. This study propounded a model called LAGProg, which used a conditional VAE as a generative model to augment the features. Survival predictions were made based on GCNN and a Cox proportional risk network. The data were taken from TCGA and included 15 datasets, none of them related to ovarian cancer. It is relevant to mention that this study found 13 prognostic markers related to breast cancer. No similar studies focused on ovarian cancer survival prediction were found. Table 1 summarizes the related studies and highlights the gap that this work aimed to address.

Table 1. Related studies in cancer survival prediction using multiomics data.

Studies	Multiomics data use	Dimensionality reduction	Graph-based interactions	ML ^a /DL ^b model	Ovarian cancer focus
Wang et al [32]	Yes	No	Yes	GCNN ^c	Yes
Jiang et al [15]	Yes	Yes (VAE ^d)	No	VAE	Yes
Zhang et al [29]	Yes	Yes (PCT ^e)	No	ML	Yes
Hira et al [35]	Yes	Yes (MMD-VAE ^f)	No	ANN ^g	Yes
Mahmoud et al [20]	Yes	Yes (VAE)	Yes	GCNN+LSTM ^h	No
Zhang et al [47]	Yes	Yes (VAE)	Yes	GCNN	No
Our study	Yes	Yes (VAE)	Proposed GCNN	VAE	Yes

^aML: machine learning.

^bDL: deep learning.

^cGCNN: graph convolutional neural network.

^dVAE: variational autoencoder.

^ePCT: principal component transformation.

^fMMD-VAE: maximum mean discrepancy VAE.

^gANN: artificial neural network.

^hLSTM: long short-term memory.

Research Gap

A review of current ovarian cancer and multiomics survival modeling literature revealed that existing studies typically address these methodological challenges in isolation. Specifically, (1) limited sample sizes continue to constrain the development of robust deep learning models for multiomics survival analysis, whereas the potential of generative models for representation learning remains underexplored; (2) high dimensionality and cross-modality heterogeneity are often managed through feature selection or linear dimensionality reduction methods rather than through deep nonlinear representation learning; (3) biological interaction modeling using GCNNs is generally explored in single-omics or pathway-focused settings rather than as part of integrated multiomics survival analysis frameworks; and (4) survival risk group identification from latent multiomics representations is rarely integrated with deep generative representation learning within a unified analytical framework.

To our knowledge, few previous studies on ovarian cancer have proposed an integrated framework that combines VAE-based nonlinear representation learning, graph-based interaction modeling, and survival stratification within a unified analytical pipeline. The present study developed and evaluated a VAE-based framework for multiomics survival stratification and introduced a graph-based extension as a biologically motivated direction for future research.

Research Objectives

This study aimed to identify prognostically distinct patient subgroups in ovarian cancer through multiomics integration and deep representation learning. To accomplish this, we pursued the following objectives:

- Mitigate the challenges associated with high-dimensional, heterogeneous multiomics data through VAE-based nonlinear representation learning
- Compress 99,322 molecular features into a compact latent representation suitable for downstream survival modeling
- Propose the integration of GCNNs to model relational biological information as a future extension of the framework

- Identify survival-associated patient subgroups through unsupervised k-means clustering of latent representations, followed by CoxPH modeling and Kaplan-Meier (KM) survival analysis to evaluate prognostic relevance

Methods

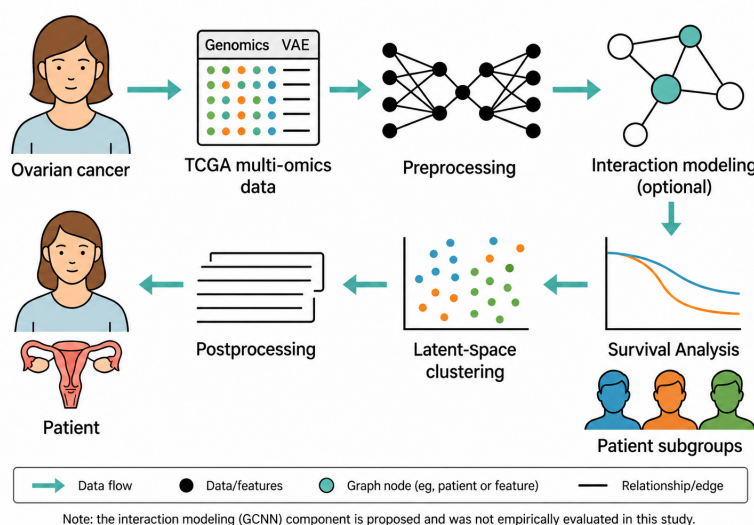
Ethical Considerations

This study is a secondary analysis of publicly available, deidentified data from TCGA accessed via the University of California, Santa Cruz (UCSC), Xena platform [48,49]. No primary data collection, participant recruitment, intervention, or direct contact with human participants was performed, and the researchers did not have access to information permitting individual identification. The Research Ethics and Scientific Integrity Office of the Pontificia Universidad Católica del Perú reviewed the project and determined that it did not require review by any of the university's research ethics committees because it did not qualify as research involving human participants, animals, or ecosystems (certificate 018-2026-NR/OETIIC/PUCP; August 28, 2026) [50]. TCGA data were accessed and used in accordance with the applicable TCGA human participant protection and data access policies [48].

Pipeline

Our analytical pipeline comprises 5 sequential stages, spanning raw multiomics preprocessing, latent representation learning, survival-based patient stratification, and model evaluation (Figure 1). The pipeline includes TCGA multiomics preprocessing; VAE-based nonlinear dimensionality reduction and latent-space representation learning; an optional GCNN component proposed for modeling interaction-aware representations but not empirically evaluated in this study; and downstream survival stratification through k-means clustering, CoxPH modeling, and KM survival analysis. The main components of the 5-stage analytical pipeline are summarized in Textbox 1.

Figure 1. Overview of the proposed multiomics analysis pipeline. GCNN: graph convolutional neural network; TCGA: The Cancer Genome Atlas; VAE: variational autoencoder.



Textbox 1. Components of the 5-stage pipeline.

Stage 1: data acquisition and preprocessing

- Retrieval of The Cancer Genome Atlas ovarian cancer multiomics data, including genomic, transcriptomic, and epigenomic layers [35]
- Harmonization of patient identifiers and integration of the omics and clinical survival datasets
- Handling of missing values using established imputation strategies [51]
- Detection and treatment of outliers following standard preprocessing practices [51]
- Featurewise normalization to harmonize measurement scales across omics modalities [29]

Stage 2: latent representation learning via variational autoencoder (VAE)

- Integration of the multiomics modalities into a unified patient-level feature matrix
- Dimensionality reduction of the combined 99,322 input features into a compact latent representation using a VAE
- Optimization of the VAE using reconstruction loss and Kullback-Leibler divergence regularization
- Extraction of patient-level latent representations for downstream clustering and survival analysis

Stage 3: proposed graph-based extension

- Conceptual construction of a graph structure representing patient-level similarity or feature-level biological associations
- Proposed definition of an adjacency matrix to encode structural relationships among graph nodes
- Proposed application of a graph convolutional neural network to refine the VAE-derived latent representations by incorporating graph-structured relational information
- This graph-based component was not implemented or empirically evaluated in the present study; consequently, all reported downstream analyses were conducted exclusively using the VAE-derived latent representations

Stage 4: survival-based patient stratification

- Unsupervised k-means clustering of the VAE-derived patient-level latent representations
- Evaluation of candidate clustering solutions using silhouette analysis, with $k=2$ selected as the optimal solution
- Cox proportional hazards (CoxPH) modeling to assess the association between cluster membership and overall survival
- Kaplan-Meier (KM) estimation and log-rank testing to evaluate differences in overall survival between the 2 identified patient subgroups

Stage 5: model evaluation

- Silhouette score assessment to quantify cluster compactness and separation; the selected 2-cluster solution achieved a silhouette score of 0.272
- Assessment of VAE reconstruction performance to evaluate the fidelity of the learned latent representation
- Evaluation of the CoxPH hazard ratio, CI, statistical significance, and proportional hazards assumption
- KM survival curve comparison and log-rank testing to evaluate the prognostic relevance of the identified patient subgroups

Data Sources and Multiomics Integration

We analyzed ovarian cancer multiomics data from TCGA obtained via the UCSC Xena data portal. The dataset included three omics layers: (1) genomics (copy number variation), (2) transcriptomics (RNA sequencing and gene expression arrays), and (3) epigenomics (DNA methylation). We downloaded sample-level matrices, where rows correspond to molecular features and columns correspond to patient samples.

The downstream analyses were conducted on the corrected full matched cohort of 291 patients (n=241, 82.8% events [deaths]; n=50, 17.2% censored). An initial barcode-matching error resulted in an incorrectly reduced cohort of 49 patients: clinical files use participant-level barcodes (12-character prefix), whereas molecular files use aliquot-level barcodes; the merge was incorrectly performed on the full string rather than the prefix, causing most matches to fail silently. After correction, 291 patients were retained. The data processing scripts and materials supporting reproducibility are available in the public GitHub repository [52].

Sample Attrition and Cohort Harmonization

A complete account of sample attrition is provided to ensure transparency and reproducibility. The Cancer Genome Atlas Ovarian Cancer (TCGA-OV) initial dataset comprised 292 patients with matched multiomics molecular profiles, corresponding to 99,322 molecular features across the selected omics layers. Following cohort harmonization, of the 292 patients, 1 (0.3%) was excluded because a valid participant-level match between the clinical and molecular datasets could not be established. Consequently, the final matched cohort consisted of 291 patients.

Inclusion criteria were as follows: (1) availability of DNA methylation (HumanMethylation450 array), RNA sequencing (Illumina HiSeq), and copy number variation (SNP6 array) data; and (2) availability of overall survival time and event status.

Exclusion criteria were as follows: (1) missing data in any omics layer after quality filtering, (2) overall survival time of 0 or less, and (3) inability to establish a valid participant-level match between the clinical and molecular datasets.

Preprocessing, Normalization, and Quality Control

Prior to model training, we applied a standard preprocessing pipeline. Features with excessive missingness were removed, and residual missing values were imputed using simple statistics (mean or median) within each omics layer. All features were subsequently standardized to zero mean and unit variance using training set statistics to mitigate scale differences between omics types and to stabilize optimization.

VAE for Nonlinear Dimensionality Reduction and Representation Learning

To address the high dimensionality and limited sample size of the integrated dataset, we used a VAE as a nonlinear latent representation learning model. Let $x \in R^D$ denote the concatenated multiomics profile for a patient. The VAE consists of an encoder network $q_\phi(z | x)$ and a decoder network $p_\theta(x | z)$:

$$q_\phi(z | x) = \mathcal{N}(\mu_\phi(x), \text{diag}(\sigma_\phi^2(x))), \quad (1)$$

$$p_\theta(x | z) = \mathcal{N}(f_\theta(z), \sigma^2 \mathbf{I}), \quad (2)$$

In these equations, $z \in \mathbb{R}^{d_z}$ is a low-dimensional latent variable, and $f_\theta(\cdot)$ denotes the decoder network. The VAE was trained by minimizing the following:

$$\mathcal{L}_{\text{VAE}}(x) = \|x - \hat{x}\|_2^2 - \beta \mathbb{E}_{q_\phi(z | x)}[\log p(z) - \log q_\phi(z | x)] \quad (3)$$

$$\mathcal{L}_{\text{recon}} = \|x - \hat{x}\|_2^2$$

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_{q_\phi(z | x)}[\log p(z) - \log q_\phi(z | x)]$$

We used the Adam optimizer with a learning rate of 10^{-3} and mini batches of size 32.

After training, the encoder was used to map each patient to a deterministic latent embedding by computing the posterior mean $\mu_\phi(x_i)$ directly without sampling. This pipeline does not use data augmentation. No synthetic patient profiles are generated at any stage. All 291 real TCGA-OV patients are used for VAE training, latent embedding extraction, k-means clustering, CoxPH modeling, and KM estimation. No train-test split is applied before VAE training. The VAE's reparameterization trick (sampling $z \sim \mathcal{N}(\mu_\phi, \text{diag}(\sigma_\phi^2))$) is used only during the forward pass for training; at inference time, the deterministic posterior mean $\mu_\phi(x_i)$ is used as each patient's embedding to ensure reproducible, noise-free representations.

GCNN

Proposed Graph-Based Extension

To explicitly capture molecular interactions and cross-omics relationships, the integrated multiomics data were modeled as a graph. Nodes represented molecular features (eg, genes or methylation probes), whereas edges encoded prior biological relationships, including coexpression, pathway comembership, or correlation thresholds computed from the training data. This process generated an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, where N denotes the number of selected molecular features represented in the graph.

Methodological Note

Although a GCNN architecture is described as a biologically motivated extension for modeling molecular interactions and cross-omics relationships, the experiments reported in this paper were performed using the VAE latent representations alone. Consequently, the GCNN formulation should be regarded as a proposed extension of the framework rather than a validated component of the present pipeline. Future work will evaluate its contribution through dedicated ablation studies and comparative analyses on larger cohorts.

In the proposed framework, a GCNN would learn interaction-aware feature representations. Given the initial node feature matrix $\mathbf{H}^{(0)}$, derived from either the VAE latent representation or the selected omics features, each GCNN layer would update the node embeddings according to the following equation:

$$\mathbf{H}^{(l+1)} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right), \quad (4)$$

where

$$\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$$

is the adjacency matrix with self-loops,

$$\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$$

is the corresponding degree matrix, $\mathbf{W}^{(l)}$ is the trainable weight matrix for the l th graph convolution layer, and $\sigma(\cdot)$ denotes a nonlinear activation function (rectified linear unit in this study). By stacking multiple GCNN layers, information would be propagated across neighboring nodes, enabling the model to capture both local and higher-order molecular interactions while integrating complementary information across multiple omics modalities.

Survival Modeling and Risk Grouping

Clinical survival data, including overall survival time and event status, were linked to each patient. CoxPH models

were used to assess the association between the learned latent representations and survival outcomes. To derive clinically interpretable survival subgroups, k-means clustering was applied to the latent space, with the optimal number of clusters selected based on silhouette analysis. Patients were subsequently assigned to one of the identified latent clusters. KM survival curves were estimated for each cluster, and the log-rank test was used to evaluate the statistical significance of differences in overall survival between the resulting patient groups.

Implementation Details

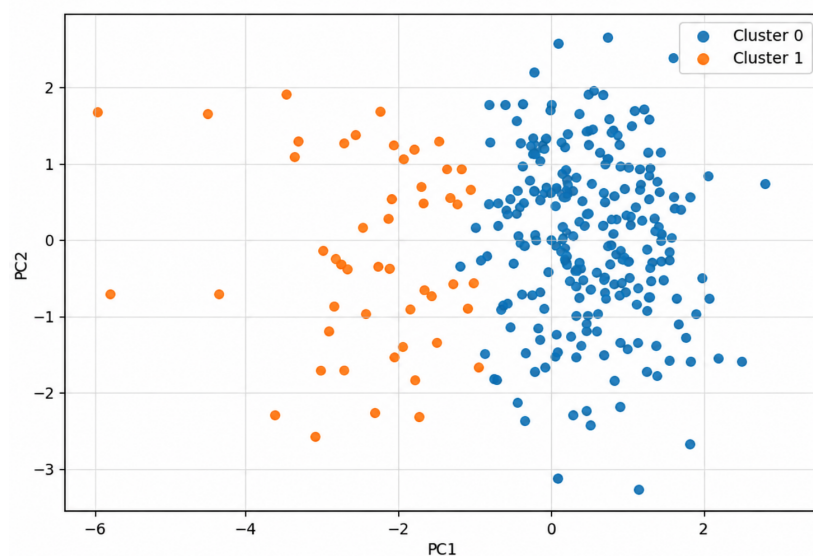
All analyses were implemented using Python (Python Software Foundation) in a reproducible Google Colab environment. Data handling and preprocessing used pandas and scikit-learn (Google Summer of Code project); the VAE model was implemented in PyTorch (Meta AI). Clustering and silhouette scores were obtained using scikit-learn, whereas CoxPH and KM analyses were conducted using standard survival analysis libraries. The complete code and configuration files are available at the UCSC Xena platform [52].

Results

Latent Representation of Multiomics Data

The refined VAE was trained using the final matched cohort of 291 TCGA-OV patients over 50 training epochs. The model learned an 8D latent representation for each patient, providing a compact embedding of the integrated multiomics data. Prior to clustering, the latent features were standardized to zero mean and unit variance to ensure equal contribution of each latent dimension to the clustering process. Figure 2 shows a 2D principal component analysis projection of the resulting latent space.

Figure 2. Principal component analysis projection of the variational autoencoder–derived latent representations showing the 2 patient clusters identified through k-means clustering ($k=2$; silhouette score=0.272). PC1: principal component 1; PC2: principal component 2.

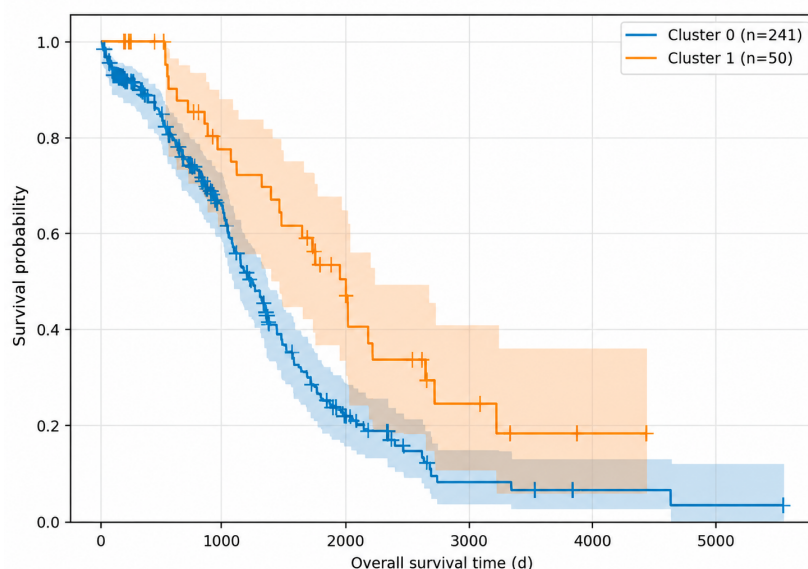


Survival Modeling With CoxPH

To quantify the prognostic significance of the latent clusters, a CoxPH model was fitted using cluster assignment as the predictor and overall survival as the outcome. The estimated hazard ratio (HR) was 0.519 (95% CI 0.343–0.785; $P=.002$), indicating that patients assigned to cluster 1 exhibited an approximately 48% lower hazard of death than those in

cluster 0. This finding demonstrates a statistically significant association between the latent cluster assignment and overall survival. The proportional hazards assumption was assessed using Schoenfeld residuals. The global test was not statistically significant ($P\approx.28$), and no systematic temporal trends were observed, supporting the validity of the CoxPH model (Figure 3).

Figure 3. Kaplan-Meier survival curves stratified by the 2 latent clusters identified from the variational autoencoder latent representations. Cluster 1 ($n=50$) exhibited significantly improved overall survival compared with cluster 0 ($n=241$). Shaded regions represent the 95% CIs, and check marks denote censored observations. The difference between survival curves was statistically significant according to the log-rank test ($P=.002$).



Unsupervised Clustering and Survival Stratification

The optimal number of clusters was determined by evaluating candidate solutions with $k=2$ to $k=8$ using silhouette analysis. The resulting silhouette scores were 0.272, 0.110,

0.140, 0.154, 0.162, 0.156, and 0.146, respectively. The highest silhouette score (0.272) was obtained for $k=2$, which was therefore selected as the optimal clustering solution. The resulting k-means clustering partitioned the cohort into cluster 0 ($n=241$) and cluster 1 ($n=50$). Figure 3 shows the KM survival curves for the 2 latent clusters. Patients

assigned to cluster 1 exhibited significantly better overall survival than those in cluster 0 ($P=.002$). The log-rank test demonstrated a statistically significant difference between the survival distributions ($\chi^2=10.0$; $P=.002$), confirming that the VAE-derived latent clusters capture prognostically distinct patient groups.

Discussion

Principal Results

We analyzed three major molecular layers—genomics, transcriptomics, and epigenomics—that together capture both stable and dynamic aspects of tumor biology. After preprocessing and harmonization, the multiomics matrices were integrated into a unified representation using a VAE, which compressed the high-dimensional feature space into a lower-dimensional latent manifold while preserving major sources of biological variation.

The pipeline also proposes an optional GCNN component to incorporate relational structure among molecular features. The rationale for the VAE and GCNN combination lies in complementary limitations: the VAE treats molecular features as conditionally independent given the latent code, discarding relational structure such as gene coexpression patterns, whereas the GCNN explicitly propagates information along edges of the molecular graph. However, this theoretical complementarity was not empirically validated in the present study. The GCNN is best understood as a theoretically motivated proposed extension whose empirical utility remains to be demonstrated.

The revised analyses demonstrate that the VAE pipeline applied to the full cohort of 291 patients identified two statistically distinct patient subgroups with significantly different overall survival (log-rank $P=.002$; HR 0.519, 95% CI 0.343–0.785; Cox model $P=.002$). Silhouette-guided cluster selection identified $k=2$ as the optimal solution (silhouette score=0.272). The proportional hazards assumption was verified using Schoenfeld residuals (global test $P\approx.28$), supporting the validity of the CoxPH model. These findings demonstrate that the VAE-derived latent representations capture prognostically distinct patient groups within the TCGA ovarian cancer cohort.

Taken together, these results demonstrate that the proposed VAE pipeline identifies statistically significant prognostic subgroups when applied to the full TCGA-OV cohort of 291 patients. The combination of silhouette-guided cluster selection, KM survival analysis, and CoxPH modeling consistently supports the prognostic relevance of the VAE-derived latent representations. These findings provide evidence that unsupervised representation learning can identify clinically meaningful patient subgroups from integrated multiomics data and establish a foundation for future extensions incorporating graph-based learning and biological pathway analysis.

Comparison With Prior Work

Our findings are consistent with those of previous studies demonstrating the utility of deep learning and VAEs for modeling the complex structure of multiomics data in cancer prognosis. Jiang et al [15] also investigated ovarian cancer using TCGA-derived multiomics data and reported that a VAE-based framework effectively captured complex molecular patterns through latent feature learning and data augmentation. Although the methodological approaches differ, the study by Jiang et al [15] and the present study support the value of representation learning for integrating heterogeneous multiomics data and improving prognostic stratification in ovarian cancer. The present study further extends this evidence by demonstrating that VAE-derived latent representations identify statistically significant prognostic subgroups within the full matched TCGA-OV cohort of 291 patients.

Limitations

First, the patient matching procedure between the clinical and molecular datasets was revised to ensure correct harmonization of TCGA participant identifiers. Consequently, all analyses in this revised manuscript were performed using the full matched TCGA-OV cohort of 291 patients ($n=241$, 82.8% events and $n=50$, 17.2% censored observations), replacing the previously analyzed subset of 49 patients. This correction substantially increases statistical power and strengthens the reliability of the reported findings.

Second, no formal ablation study was conducted to quantify the individual contribution of each component of the proposed pipeline. Specifically, the survival stratification performance was not compared across (1) raw multiomics features, (2) VAE-derived latent embeddings, and (3) VAE embeddings augmented with graph-based learning. Consequently, the incremental benefit of each component, particularly the proposed GCNN extension, could not be independently quantified.

Third, the VAE was trained in an unsupervised manner without incorporating clinical outcome information during representation learning. Future work may investigate supervised or semisupervised approaches that integrate survival objectives into the latent-space optimization or incorporate molecular interaction networks directly within the representation learning process.

Future Work

Future research should extend the present framework in several directions. First, external validation using independent multiomics ovarian cancer cohorts will be essential to assess the generalizability and clinical applicability of the proposed methodology. Second, formal ablation studies should be conducted to quantify the contribution of each component of the analysis pipeline, including comparisons among raw multiomics features, VAE-derived latent representations, and future graph-based extensions. Third, supervised or semisupervised representation learning approaches, such as survival-informed VAEs, may further improve prognostic

stratification by incorporating clinical outcome information during latent-space optimization. Finally, future work may integrate GNN architectures and biologically informed gene-gene or patient-patient similarity networks to better capture molecular interactions and enhance both predictive performance and model interpretability.

Conclusions

This study demonstrates that a VAE-based framework can integrate multiomics data to identify prognostically distinct patient subgroups in ovarian cancer. Applied to a matched TCGA-OV cohort of 291 patients, the proposed pipeline identified two latent clusters with significantly different overall survival, as demonstrated through KM analysis

(log-rank $P=.002$) and CoxPH modeling (HR 0.519, 95% CI 0.343-0.785; $P=.002$). The optimal two-cluster solution was supported by silhouette analysis, and the proportional hazards assumption was satisfied, supporting the validity of the CoxPH model. These findings demonstrate the potential of unsupervised deep representation learning for survival stratification using integrated multiomics data. Future work should focus on external validation in independent cohorts, comprehensive biological characterization of the identified latent clusters, formal ablation studies to quantify the contribution of individual pipeline components, and the evaluation of graph-based learning approaches to further enhance predictive performance and interpretability.

Acknowledgments

Generative AI tools were not used in any portion of writing, data analysis, code development, or figure generation of this manuscript. All text was written by the human authors.

Funding

The authors declared no financial support was received for this work.

Data Availability

The data analyzed in this study are publicly available through The Cancer Genome Atlas and were accessed through the University of California, Santa Cruz, Xena platform [52]. The code, data processing scripts, model configurations, and materials supporting reproducibility of the analyses are publicly available in the authors' GitHub repository [49]. No new identifiable participant data were collected or generated in this study.

Authors' Contributions

Conceptualization: CM, CDP

Data curation: CM

Methodology: CM, CDP

Supervision: CM

Writing—original draft: CM, CDP

Writing—review and editing: CM, CDP

Conflicts of Interest

None declared.

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024;74(3):229-263. [doi: [10.3322/caac.21834](https://doi.org/10.3322/caac.21834)] [Medline: [38572751](https://pubmed.ncbi.nlm.nih.gov/38572751/)]
2. Li J, Kuang X. Global cancer statistics of young adults and its changes in the past decade: incidence and mortality from GLOBOCAN 2022. *Public Health*. Dec 2024;237:336-343. [doi: [10.1016/j.puhe.2024.10.033](https://doi.org/10.1016/j.puhe.2024.10.033)] [Medline: [39515218](https://pubmed.ncbi.nlm.nih.gov/39515218/)]
3. Hira MT, Razzaque MA, Sarker M. Ovarian cancer data analysis using deep learning: a systematic review. *Eng Appl Artif Intell*. Dec 2024;138:109250. [doi: [10.1016/j.engappai.2024.109250](https://doi.org/10.1016/j.engappai.2024.109250)]
4. Xiao Y, Bi M, Guo H, Li M. Multi-omics approaches for biomarker discovery in early ovarian cancer diagnosis. *EBioMedicine*. May 2022;79:104001. [doi: [10.1016/j.ebiom.2022.104001](https://doi.org/10.1016/j.ebiom.2022.104001)] [Medline: [35439677](https://pubmed.ncbi.nlm.nih.gov/35439677/)]
5. Li X, Budzin A, Wang Y. Editorial: Application and innovation of multiomics technologies in clinical oncology. *Front Oncol*. Mar 2023;13:1179829. [doi: [10.3389/fonc.2023.1179829](https://doi.org/10.3389/fonc.2023.1179829)] [Medline: [37056350](https://pubmed.ncbi.nlm.nih.gov/37056350/)]
6. Yelmen B, Jay F. An overview of deep generative models in functional and evolutionary genomics. *Annu Rev Biomed Data Sci*. Aug 10, 2023;6:173-189. [doi: [10.1146/annurev-biodatasci-020722-115651](https://doi.org/10.1146/annurev-biodatasci-020722-115651)] [Medline: [37137168](https://pubmed.ncbi.nlm.nih.gov/37137168/)]
7. Ramadan A, Jarab AS, Al Meslamani AZ, Alzoubi KH. Hurdles in the personalized medicine implementation path: public's literacy and misconceptions of whole genomic sequencing test. *Crit Public Health*. Dec 31, 2025;35(1). [doi: [10.1080/09581596.2025.2488119](https://doi.org/10.1080/09581596.2025.2488119)]
8. Vogeser M, Bendt AK. From research cohorts to the patient - a role for "omics" in diagnostics and laboratory medicine? *Clin Chem Lab Med*. 2023;61(6):974-980. [doi: [10.1515/cclm-2022-1147](https://doi.org/10.1515/cclm-2022-1147)] [Medline: [36592431](https://pubmed.ncbi.nlm.nih.gov/36592431/)]

9. Vlachavas EI, Bohn J, Ückert F, Nürnberg S. A detailed catalogue of multi-omics methodologies for identification of putative biomarkers and causal molecular networks in translational cancer research. *Int J Mol Sci.* Mar 10, 2021;22(6):2822. [doi: [10.3390/ijms22062822](https://doi.org/10.3390/ijms22062822)] [Medline: [33802234](https://pubmed.ncbi.nlm.nih.gov/33802234/)]
10. Sigman M. Introduction: personalized medicine: what is it and what are the challenges? *Fertil Steril.* Jun 2018;109(6):944-945. [doi: [10.1016/j.fertnstert.2018.04.027](https://doi.org/10.1016/j.fertnstert.2018.04.027)] [Medline: [29935651](https://pubmed.ncbi.nlm.nih.gov/29935651/)]
11. Niederberger C. Re: Introduction: personalized medicine: what is it and what are the challenges?. *J Urol.* Mar 2019;201(3):426-427. [doi: [10.1097/01.JU.0000553669.64428.19](https://doi.org/10.1097/01.JU.0000553669.64428.19)] [Medline: [30759644](https://pubmed.ncbi.nlm.nih.gov/30759644/)]
12. Schleidgen S, Klingler C, Bertram T, Rogowski WH, Marckmann G. What is personalized medicine: sharpening a vague term based on a systematic literature review. *BMC Med Ethics.* Dec 21, 2013;14:55. [doi: [10.1186/1472-6939-14-55](https://doi.org/10.1186/1472-6939-14-55)] [Medline: [24359531](https://pubmed.ncbi.nlm.nih.gov/24359531/)]
13. Ghebrehwet I, Zaki N, Damseh R, Mohamad MS. Revolutionizing personalized medicine with generative AI: a systematic review. *Artif Intell Rev.* 2024;57(128). [doi: [10.1007/s10462-024-10768-5](https://doi.org/10.1007/s10462-024-10768-5)]
14. Velmurugan S, Wankhar D, Paramasivan V, Subbaraj GK. Technological innovations and multi-omics approaches in cancer research: a comprehensive review. *BIOCELL.* 2025;49(8):1363-1390. [doi: [10.32604/biocell.2025.065891](https://doi.org/10.32604/biocell.2025.065891)]
15. Jiang L, Xu C, Bai Y, et al. Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data. *NPJ Precis Oncol.* Jan 5, 2024;8(1):4. [doi: [10.1038/s41698-023-00494-6](https://doi.org/10.1038/s41698-023-00494-6)] [Medline: [38182734](https://pubmed.ncbi.nlm.nih.gov/38182734/)]
16. Gu J, Wu Y, Tao W, et al. An integrative multi-omics study to identify candidate DNA methylation biomarkers associated with gastric cancer prognosis. *Arch Toxicol.* Oct 2025;99(10):4067-4080. [doi: [10.1007/s00204-025-04118-9](https://doi.org/10.1007/s00204-025-04118-9)] [Medline: [40616607](https://pubmed.ncbi.nlm.nih.gov/40616607/)]
17. Lin J, Wang J, Zhao K, Li Y, Zhang X, Sheng J. Molecular targets and mechanisms of traditional Chinese medicine combined with chemotherapy for gastric cancer: a meta-analysis and multi-omics approach. *Ann Med.* Dec 2025;57(1):2494671. [doi: [10.1080/07853890.2025.2494671](https://doi.org/10.1080/07853890.2025.2494671)] [Medline: [40317214](https://pubmed.ncbi.nlm.nih.gov/40317214/)]
18. Ge J, Cai J, Zhang G, Li D, Tao L. Multi-omics integration and machine learning uncover molecular basal-like subtype of pancreatic cancer and implicate A2ML1 in promoting tumor epithelial-mesenchymal transition. *J Transl Med.* Jul 4, 2025;23(1):741. [doi: [10.1186/s12967-025-06711-z](https://doi.org/10.1186/s12967-025-06711-z)] [Medline: [40615919](https://pubmed.ncbi.nlm.nih.gov/40615919/)]
19. Khella CA, Mehta GA, Mehta RN, Gatz ML. Recent advances in integrative multi-omics research in breast and ovarian cancer. *J Pers Med.* Feb 19, 2021;11(2):149. [doi: [10.3390/jpm11020149](https://doi.org/10.3390/jpm11020149)] [Medline: [33669749](https://pubmed.ncbi.nlm.nih.gov/33669749/)]
20. Mahmoud A, Alhussein M, Aurangzeb K, Takaoka E. Breast cancer survival prediction modeling based on genomic data: an improved prognosis-driven deep learning approach. *IEEE Access.* 2024;12:119502-119519. [doi: [10.1109/ACCESS.2024.3449814](https://doi.org/10.1109/ACCESS.2024.3449814)]
21. Zhong Z, Wang Z, Fan T, Xu L, Li Q, Dong C. Multi-omics analysis reveals the impact of thrombotic diseases on the occurrence and prognosis of breast cancer. *J Steroid Biochem Mol Biol.* Oct 2025;253:106815. [doi: [10.1016/j.jsbmb.2025.106815](https://doi.org/10.1016/j.jsbmb.2025.106815)] [Medline: [40513964](https://pubmed.ncbi.nlm.nih.gov/40513964/)]
22. Omran MM, Emam M, Gamaleldin M, Abushady AM, Elattar MA, El-Hadidi M. Comparative analysis of statistical and deep learning-based multi-omics integration for breast cancer subtype classification. *J Transl Med.* Jul 1, 2025;23(1):709. [doi: [10.1186/s12967-025-06662-5](https://doi.org/10.1186/s12967-025-06662-5)] [Medline: [40598554](https://pubmed.ncbi.nlm.nih.gov/40598554/)]
23. Zhang Y, Chen F, Creighton CJ. Pan-cancer, multi-omic correlates of survival transcending tumor lineage across 11,019 patients reveal targets and pathways. *NPJ Precis Onc.* Jul 2025;9(1):226. [doi: [10.1038/s41698-025-01029-x](https://doi.org/10.1038/s41698-025-01029-x)] [Medline: [40617938](https://pubmed.ncbi.nlm.nih.gov/40617938/)]
24. Wang C, Zhang M, Zhao J, Li B, Xiao X, Zhang Y. The prediction of drug sensitivity by multi-omics fusion reveals the heterogeneity of drug response in pan-cancer. *Comput Biol Med.* Sep 2023;163:107220. [doi: [10.1016/j.compbimed.2023.107220](https://doi.org/10.1016/j.compbimed.2023.107220)] [Medline: [37406589](https://pubmed.ncbi.nlm.nih.gov/37406589/)]
25. Li Y, Chen C, Ji X, et al. Multi-omics analysis and experiments uncover the link between cancer intrinsic drivers, stemness, and immunotherapy in ovarian cancer with validation in a pan-cancer census. *Front Immunol.* May 2025;16:1549656. [doi: [10.3389/fimmu.2025.1549656](https://doi.org/10.3389/fimmu.2025.1549656)] [Medline: [40406120](https://pubmed.ncbi.nlm.nih.gov/40406120/)]
26. Loizzi V, Comes MC, Arezzo F, et al. Validation of machine learning-based models to predict and explain the risk of ovarian cancer: a multicentric study on *BRCA*-mutated patients undergoing risk-reducing salpingo-oophorectomy. *Front Oncol.* 2025;15:1574037. [doi: [10.3389/fonc.2025.1574037](https://doi.org/10.3389/fonc.2025.1574037)] [Medline: [40303993](https://pubmed.ncbi.nlm.nih.gov/40303993/)]
27. Kliuchnikova A, Gordeeva A, Abdurakhimov A, et al. Ovarian cancer: multi-omics data integration. *Int J Mol Sci.* Jun 21, 2025;26(13):5961. [doi: [10.3390/ijms26135961](https://doi.org/10.3390/ijms26135961)] [Medline: [40649740](https://pubmed.ncbi.nlm.nih.gov/40649740/)]
28. Alharbi F, Vakanski A, Zhang B, Elbashir MK, Mohammed M. Comparative analysis of multi-omics integration using graph neural networks for cancer classification. *IEEE Access.* 2025;13:37724-37736. [doi: [10.1109/access.2025.3540769](https://doi.org/10.1109/access.2025.3540769)] [Medline: [40123934](https://pubmed.ncbi.nlm.nih.gov/40123934/)]

29. Zhang Z, Wei Z, Zhao L, Gu C, Meng Y. Assessing the clinical utility of multi-omics data for predicting serous ovarian cancer prognosis. *J Obstet Gynaecol*. Dec 2023;43(1):2171778. [doi: [10.1080/01443615.2023.2171778](https://doi.org/10.1080/01443615.2023.2171778)] [Medline: [36803381](https://pubmed.ncbi.nlm.nih.gov/36803381/)]
30. Asadi F, Rahimi M, Ramezanghorbani N, Almasi S. Comparing the effectiveness of artificial intelligence models in predicting ovarian cancer survival: a systematic review. *Cancer Rep (Hoboken)*. Mar 2025;8(3):e70138. [doi: [10.1002/cmr2.70138](https://doi.org/10.1002/cmr2.70138)] [Medline: [40103563](https://pubmed.ncbi.nlm.nih.gov/40103563/)]
31. Braytee A, He S, Tang S, et al. Identification of cancer risk groups through multi-omics integration using autoencoder and tensor analysis. *Sci Rep*. May 17, 2024;14(1):11263. [doi: [10.1038/s41598-024-59670-8](https://doi.org/10.1038/s41598-024-59670-8)] [Medline: [38760420](https://pubmed.ncbi.nlm.nih.gov/38760420/)]
32. Wang H, Han X, Niu S, Cheng H, Ren J, Duan Y. DFASGCNS: a prognostic model for ovarian cancer prediction based on dual fusion channels and stacked graph convolution. *PLoS One*. Dec 2024;19(12):e0315924. [doi: [10.1371/journal.pone.0315924](https://doi.org/10.1371/journal.pone.0315924)] [Medline: [39680618](https://pubmed.ncbi.nlm.nih.gov/39680618/)]
33. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med*. Sep 27, 2021;13(1):152. [doi: [10.1186/s13073-021-00968-x](https://doi.org/10.1186/s13073-021-00968-x)] [Medline: [34579788](https://pubmed.ncbi.nlm.nih.gov/34579788/)]
34. Ghantasala GS, Dilip K, Vidyullatha P, et al. Enhanced ovarian cancer survival prediction using temporal analysis and graph neural networks. *BMC Med Inform Decis Mak*. Oct 10, 2024;24(1):299. [doi: [10.1186/s12911-024-02665-2](https://doi.org/10.1186/s12911-024-02665-2)] [Medline: [39390514](https://pubmed.ncbi.nlm.nih.gov/39390514/)]
35. Hira MT, Razzaque MA, Angione C, Scrivens J, Sawan S, Sarker M. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. *Sci Rep*. Mar 18, 2021;11(1):6265. [doi: [10.1038/s41598-021-85285-4](https://doi.org/10.1038/s41598-021-85285-4)] [Medline: [33737557](https://pubmed.ncbi.nlm.nih.gov/33737557/)]
36. Al-khassaweneh M, Bronakowski M, Al-Sharoa E. Multivariate and dimensionality-reduction-based machine learning techniques for tumor classification of RNA-Seq data. *Appl Sci*. 2023;13(23):12801. [doi: [10.3390/app132312801](https://doi.org/10.3390/app132312801)]
37. Kalantan ZI, Alqarni LZ, Binhim SM. Unveiling hidden insights: dimensionality reduction for prostate cancer data with PCA and Gaussian mixture model. *Adv Appl Stat*. 2025;92(4):583-602. [doi: [10.17654/0972361725024](https://doi.org/10.17654/0972361725024)]
38. Li X, Chen X, Rezaeipana A. Automatic breast cancer diagnosis based on hybrid dimensionality reduction technique and ensemble classification. *J Cancer Res Clin Oncol*. Aug 2023;149(10):7609-7627. [doi: [10.1007/s00432-023-04699-x](https://doi.org/10.1007/s00432-023-04699-x)] [Medline: [36995408](https://pubmed.ncbi.nlm.nih.gov/36995408/)]
39. Al-Turaiki I. Dimensionality reduction of RNA-Seq data. *Int J Comput Sci Netw Secur*. Mar 2021;21(3):31-36. [doi: [10.22937/IJCSNS.2021.21.3.4](https://doi.org/10.22937/IJCSNS.2021.21.3.4)]
40. Simidjievski N, Bodnar C, Tariq I, et al. Variational autoencoders for cancer data integration: design principles and computational practice. *Front Genet*. 2019;10:1205. [doi: [10.3389/fgene.2019.01205](https://doi.org/10.3389/fgene.2019.01205)] [Medline: [31921281](https://pubmed.ncbi.nlm.nih.gov/31921281/)]
41. Li Y, Zheng D, Sun K, et al. A densely connected framework for cancer subtype classification. *BMC Bioinformatics*. Jul 18, 2025;26(1):183. [doi: [10.1186/s12859-025-06230-0](https://doi.org/10.1186/s12859-025-06230-0)] [Medline: [40681997](https://pubmed.ncbi.nlm.nih.gov/40681997/)]
42. Cui S, Luo Y, Tseng HH, Ten Haken RK, El Naqa I. Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage. *Med Phys*. May 2019;46(5):2497-2511. [doi: [10.1002/mp.13497](https://doi.org/10.1002/mp.13497)] [Medline: [30891794](https://pubmed.ncbi.nlm.nih.gov/30891794/)]
43. Yan J, Ma M, Yu Z. bmVAE: a variational autoencoder method for clustering single-cell mutation data. *Bioinformatics*. Jan 1, 2023;39(1):btac790. [doi: [10.1093/bioinformatics/btac790](https://doi.org/10.1093/bioinformatics/btac790)] [Medline: [36478203](https://pubmed.ncbi.nlm.nih.gov/36478203/)]
44. Paul SG, Saha A, Hasan MZ, Noori SR, Moustafa A. A systematic review of graph neural network in healthcare-based applications: recent advances, trends, and future directions. *IEEE Access*. 2024;12:15145-15170. [doi: [10.1109/ACCESS.2024.3354809](https://doi.org/10.1109/ACCESS.2024.3354809)]
45. Li R, Yuan X, Radfar M, et al. Graph signal processing, graph neural network and graph learning on biological data: a systematic review. *IEEE Rev Biomed Eng*. 2023;16:109-135. [doi: [10.1109/RBME.2021.3122522](https://doi.org/10.1109/RBME.2021.3122522)] [Medline: [34699368](https://pubmed.ncbi.nlm.nih.gov/34699368/)]
46. Gogoshin G, Rodin AS. Graph neural networks in cancer and oncology research: emerging and future trends. *Cancers (Basel)*. Dec 15, 2023;15(24):5858. [doi: [10.3390/cancers15245858](https://doi.org/10.3390/cancers15245858)] [Medline: [38136405](https://pubmed.ncbi.nlm.nih.gov/38136405/)]
47. Zhang Y, Xiong S, Wang Z, et al. Local augmented graph neural network for multi-omics cancer prognosis prediction and analysis. *Methods*. May 2023;213:1-9. [doi: [10.1016/j.ymeth.2023.02.011](https://doi.org/10.1016/j.ymeth.2023.02.011)] [Medline: [36933628](https://pubmed.ncbi.nlm.nih.gov/36933628/)]
48. The Cancer Genome Atlas Program: human subjects protection and data access policies. National Cancer Institute. 2014. URL: <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history/policies/tcga-human-subjects-data-policies.pdf> [Accessed 2026-08-28]
49. Goldman M. Explore TCGA, GDC, and other public cancer genomics resources. University of California, Santa Cruz. 2019. URL: <https://xena.ucsc.edu/public/> [Accessed 2026-08-28]
50. Ética de la investigación e integridad científica [Webpage in Spanish]. Vicerrectorado de Investigación, Pontificia Universidad Católica del Perú. URL: https://investigacion.pucp.edu.pe/etica-e-integridad/?utm_source=chatgpt.com [Accessed 2026-08-28]

51. Dhingra H, Shetty R. Comparative study of machine learning and deep learning models for early prediction of ovarian cancer. *IEEE Access*. 2025;13:87336-87349. [doi: [10.1109/ACCESS.2025.3567081](https://doi.org/10.1109/ACCESS.2025.3567081)]
52. TCGA-ovarian-MultiOmics-DeepLearning. GitHub. URL: <https://github.com/carloshachi777/multiomics-ovarian-survival> [Accessed 2026-09-04]

Abbreviations

CoxPH: Cox proportional hazards

DFASGCNS: Dual Fusion Channels and Stacked Graph Convolutional Neural Network

GCNN: graph convolutional neural network

GNN: graph neural network

HR: hazard ratio

KM: Kaplan-Meier

ML: machine learning

TCGA: The Cancer Genome Atlas

TCGA-OV: The Cancer Genome Atlas Ovarian Cancer

UCSC: University of California, Santa Cruz

VAE: variational autoencoder

XGBoost: Extreme Gradient Boosting

Edited by Qianqian Song; peer-reviewed by Fei Long, Minmin Yu; submitted 05.Dec.2025; final revised version received 28.Aug.2026; accepted 31.Aug.2026; published 21.Sep.2026

Please cite as:

Marino C, Diaz Paz C

Deep Generative and Graph-Based Representation Learning for Multiomics Survival Stratification in Ovarian Cancer: Secondary Analysis

JMIR Bioinform Biotech 2026;7:e89069

URL: <https://bioinform.jmir.org/2026/1/e89069>

doi: [10.2196/89069](https://doi.org/10.2196/89069)

© Carlos Marino, Claudia Diaz Paz. Originally published in JMIR Bioinformatics and Biotechnology (<https://bioinform.jmir.org>), 21.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Bioinformatics and Biotechnology, is properly cited. The complete bibliographic information, a link to the original publication on <https://bioinform.jmir.org/>, as well as this copyright and license information must be included.