Computer Methods and Programs in Biomedicine

CLM-GVAE: Contrastive Learning Multi-modal Graph based Variational Autoencoders for Predicting Immunotherapy Response

--Manuscript Draft--

Manuscript Number:	CMPB-D-25-03910
Article Type:	Full Length Article
Section/Category:	Diagnostics and Decision Support
Keywords:	Multi-modal Deep Learning; Contrastive Learning; Graph Variational Autoencoder; Immunotherapy Response Prediction; NSCLC
Corresponding Author:	Quang H. Nguyen Hanoi University of Science and Technology VIET NAM
First Author:	Anh Le
Order of Authors:	Anh Le
	Quang H. Nguyen
Manuscript Region of Origin:	Asia Pacific
Abstract:	Objective: Predicting immunotherapy response in non-small cell lung cancer (NSCLC) is hindered by the challenge of integrating heterogeneous, multi-modal patient data. Our objective was to develop a deep learning framework that explicitly aligns these data modalities to improve predictive accuracy. Methods:We propose CLM-GVAE, an end-to-end framework that uses view-specific Graph Variational Autoencoders (GVAEs) to model relational patient data. The model incorporates a hierarchical attention mechanism to aggregate features from multiple radiological lesions and employs a projected cross-view contrastive learning objective to align the latent spaces of clinical, pathology, and radiology data into a shared, synergistic embedding. Results: On a public cohort of 247 NSCLC patients, CLM-GVAE achieved a mean AUC of 0.806 ± 0.088, significantly outperforming standard machine learning baselines and ablated versions of our model. These results validate the crucial contributions of both the lesion-aware aggregation and the contrastive alignment strategy. Conclusions: Explicitly aligning latent representations from multi-modal data using a graph-based, contrastive learning approach is a robust and effective strategy for developing more accurate prognostic models for immunotherapy response.
Opposed Reviewers:	

School of Information and Communication Technology,
Hanoi University of Science and Technology
1 Dai Co Viet Road, Ha Noi, Viet Nam
Tel: (+84) 24 3869 2463

July 12th, 2025

Dear **Executive Editors of Computer Methods and Programs in Biomedicine Journal,**

We are pleased to submit our manuscript entitled "*CLM-GVAE: Contrastive Learning Multi-modal Graph based Variational Autoencoders for Predicting Immunotherapy Response in Non-Small Cell Lung Cancer*" for consideration as an original research article in **Computer Methods and Programs in Biomedicine Journal**.

We confirm that this manuscript is original, has not been published elsewhere, and is not currently under consideration by another journal. All authors have approved the manuscript and agree with its submission to Computer Methods and Programs in Biomedicine Journal. Furthermore, we confirm that our use of the dataset, originally published by Vanguri et al. (2022) and publicly available on the Synapse platform (ID: syn26642505), complies with its terms of use and is properly cited within the manuscript. To the best of our knowledge, the work is free of any plagiarism or fraudulent data.

Problem: The accurate prediction of patient response to immunotherapy in non-small cell lung cancer (NSCLC) remains a critical challenge in precision oncology, largely due to the difficulty of effectively integrating heterogeneous multi-modal data. Current state-of-the-art methods, while advancing the field, often struggle with several key limitations. They frequently fail to explicitly align the latent spaces of different modalities, resulting in suboptimal fusion of what remain disparate feature sets. Furthermore, they often oversimplify complex intra-modal structures, such as the heterogeneity among multiple radiological lesions within a single patient, by using simple aggregation techniques. These gaps, compounded by the practical challenge of extensive data missingness in real-world cohorts, hinder the development of truly synergistic prognostic models and limit their clinical applicability.

Contribution: We address these challenges by introducing CLM-GVAE, a novel end-to-end deep learning framework. Our primary contribution is a projected cross-view contrastive learning strategy that explicitly aligns the latent spaces of graph-contextualized representations from clinical, pathology, and radiology data. This alignment is complemented by two other key innovations: a hierarchical attention mechanism to produce a nuanced patient-level signature from multi-lesion radiological data, and the foundational use of Graph Variational

Autoencoders to effectively model the inherent relational structure within each data modality. Evaluated on a public NSCLC cohort, our framework demonstrates state-of-the-art performance, achieving a mean AUC of 0.806 and significantly outperforming established baselines.

We believe this work is an excellent fit for the Journal of Biomedical Informatics as it presents a novel computational methodology directly addressing a critical challenge in clinical bioinformatics. Our focus on integrating complex, real-world biomedical data and developing a robust, interpretable model for prognostic prediction aligns perfectly with the journal's scope of advancing clinical decision-making through informatics.

Sincerely,

Quang H. Nguyen, Ph.D. (On behalf of the authors)
Lecturer in Data Science
School of Information and Communication Technology
Hanoi University of Science and Technology, Vietnam

Highlights

CLM-GVAE: Contrastive Learning Multi-modal Graph based Variational Autoencoders for Predicting Immunotherapy Response

Anh K. Le, Quang H. Nguyen

- Develop CLM-GVAE, a deep learning end-to-end framework that explicitly aligns heterogeneous, multi-modal patient data.
- Propose an hierarchical attention mechanism to aggregate features from multiple radiological lesions.
- Employs a projected cross-view contrastive learning objective to align the latent spaces of clinical, pathology, and radiology data into an embedding.
- CLM-GVAE is an accurate prognostic models for immunotherapy response.

Aligning Multi-modal Patient Data to Improve Immunotherapy Prediction

Input

Clinical

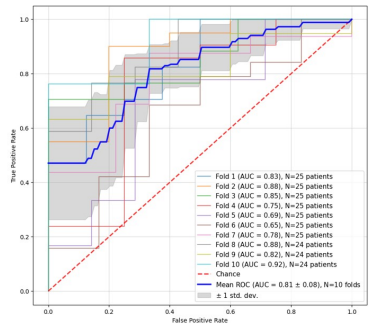
Multi-Lesion Radiology

Pathology

Method (CLM-GVAE)

CLM-GVAE

Output



Improved Immunotherapy Response Prediction

CLM-GVAE: Contrastive Learning Multi-modal Graph based Variational Autoencoders for Predicting Immunotherapy Response

Anh K. Le[a], Quang H. Nguyen[a,*]

[a]*School of Information and Communication Technology, Hanoi University of Science and Technology, 1 Dai Co Viet, Hanoi, 100000, Vietnam*

Abstract

Objective: Predicting immunotherapy response in non-small cell lung cancer (NSCLC) is hindered by the challenge of integrating heterogeneous, multi-modal patient data. Our objective was to develop a deep learning framework that explicitly aligns these data modalities to improve predictive accuracy.

Methods: We propose CLM-GVAE, an end-to-end framework that uses view-specific Graph Variational Autoencoders (GVAEs) to model relational patient data. The model incorporates a hierarchical attention mechanism to aggregate features from multiple radiological lesions and employs a projected cross-view contrastive learning objective to align the latent spaces of clinical, pathology, and radiology data into a shared, synergistic embedding.

Results: On a public cohort of 247 NSCLC patients, CLM-GVAE achieved a mean AUC of 0.806 ± 0.088, significantly outperforming standard machine learning baselines (e.g., XGBoost, AUC 0.680) and ablated versions of our model. These results validate the crucial contributions of both the lesion-aware aggregation and the contrastive alignment strategy.

Conclusions: Explicitly aligning latent representations from multi-modal data using a graph-based, contrastive learning approach is a robust and effective strategy for developing more accurate prognostic models for immunotherapy response.

Keywords: Multi-modal Deep Learning, Contrastive Learning, Graph Variational Autoencoder, Immunotherapy Response Prediction, NSCLC

*Corresponding authors.

1. Introduction

The advent of immune checkpoint inhibitors, particularly those targeting the PD-1/PD-L1 axis, has revolutionized the treatment landscape for non-small cell lung cancer (NSCLC) [1, 2]. However, a significant portion of patients do not benefit from these therapies, and identifying robust predictive biomarkers remains a critical challenge in precision oncology [3, 4]. While individual biomarkers like PD-L1 expression and Tumor Mutational Burden (TMB) offer modest predictive value, there is a growing consensus that integrating information from diverse, routinely collected clinical data modalities, including clinical records, pathology, genomics, and radiology is paramount for developing more accurate prognostic models [5, 6, 7].

Despite the promise of multi-modal integration, current computational approaches face significant hurdles. Early methods relying on simple feature concatenation or late-stage prediction fusion often fail to capture the complex, non-linear interactions between modalities [5]. While more advanced deep learning frameworks utilizing attention mechanisms, inspired by the Transformer architecture [8], have shown promise [9, 10, 11], a fundamental challenge persists: ensuring that the learned representations are not merely combined, but are semantically aligned in a shared latent space where features from different views of the same patient converge. Furthermore, a critical and often overlooked limitation is the handling of intra-modality heterogeneity. In radiology, for example, patients frequently present with multiple lesions of varying size, shape, and texture. Standard approaches that aggregate these features via simple mean or max pooling risk obscuring vital prognostic signals by failing to account for the unique contribution of each lesion. This problem is effectively a Multiple Instance Learning (MIL) challenge, where attention-based aggregators have proven to be a state-of-the-art solution [12, 13]. Finally, translating these models into clinical practice requires robust strategies for handling the pervasive issue of missing data across views.

To address the aforementioned limitations, this study introduces and evaluates **CLM-GVAE** (*Contrastive Learning Multi-modal Graph based Variational Autoencoders*), a novel end-to-end deep learning framework designed to learn robust and aligned representations from heterogeneous patient data for predicting immunotherapy response.

The specific research objectives are:

1. We develop a view-specific Graph Variational Autoencoder (GVAE) architecture that learns a probabilistic latent representation for each modality. This architecture, founded on the principles of GNNs [14, 15] and VAEs [16], processes both patient features and the relational structure of patient similarity graphs using advanced Graph Attention Networks [17, 18].
2. We introduce an attention-based lesion aggregator that learns to dynamically weigh and combine features from multiple, distinct radiological lesions into a single, comprehensive patient-level signature, directly addressing the challenge of intra-modality heterogeneity [12].
3. We employ a projected cross-view contrastive learning objective [19] to explicitly align the latent spaces of the different modalities, forcing the model to learn a shared, view-invariant semantic representation for each patient.

We evaluate CLM-GVAE on a public cohort of 247 NSCLC patients treated with immunotherapy [20]. Our framework achieves a mean AUC of 0.806, significantly outperforming traditional machine learning baselines, unimodal models, and a re-implementation of the original study's deep learning approach. Comprehensive ablation studies confirm that the cross-view contrastive alignment and the lesion aggregation mechanism are critical drivers of this performance. CLM-GVAE provides a robust, versatile, and principled framework for multi-modal data integration, demonstrating a powerful pathway toward more accurate prognostic modeling in oncology.

The remainder of this paper is organized as follows. Section 2 reviews related work in multi-modal data fusion, graph representation learning, and contrastive alignment for healthcare applications. Section 3 provides a detailed description of our proposed CLM-GVAE framework, including its core architectural components, the training process, and the composite loss function. Section 4 presents the comprehensive experimental evaluation, including comparisons with baseline models, detailed ablation studies, and a discussion of the results. Finally, Section 5 concludes the paper by summarizing our findings, acknowledging limitations, and outlining directions for future research.

2. Related Works

The challenge of predicting patient outcomes from complex medical data has spurred significant research at the intersection of multi-modal fusion, graph representation learning, and contrastive alignment. Our work builds upon foundational concepts in these areas while directly addressing their limitations in the context of clinical oncology.

2.1. Multi-modal Data Fusion in Oncology

The integration of heterogeneous data sources, such as radiology images, pathology slides, genomic assays, and clinical records, is paramount for a holistic understanding of cancer [3, 6, 7]. This process, however, come with challenges, including data heterogeneity, dimensionality, and missingness [21]. Early deep learning approaches focused on learning modality-specific representations and then combining them through either late-stage prediction fusion or simple intermediate feature concatenation [9]. More recent state-of-the-art models, such as the DyAM model proposed by Vanguri et al [20]. and other Transformer-based architectures [10, 11], have successfully employed attention mechanisms [8] to dynamically weigh the contribution of each modality. However, these methods, while powerful, often lack an explicit mechanism to enforce semantic consistency between the latent spaces of different views. This can result in a fused representation that is a well-weighted combination of disparate features rather than a truly synergistic embedding learned from a shared, aligned space. Our work addresses this gap by using contrastive learning as a core objective to enforce this alignment.

2.2. Graph-based Representation Learning in Healthcare

Patient data often possesses an inherent relational structure, which can be modeled by constructing patient similarity graphs, an approach that has proven effective for tasks like disease prediction and patient stratification [22, 23, 24]. Graph Neural Networks (GNNs), such as Graph Attention Networks (GATs) [18], are highly effective at leveraging this structure to learn context-aware patient embeddings [25, 26, 27, 28, 29]. Furthermore, Graph Variational Autoencoders (GVAEs) [15][20] combine the relational inductive bias of GNNs [14] with the generative power of VAEs [16] to learn robust, probabilistic latent distributions for graph-structured data. However, while GVAEs are excellent for unimodal representation learning, their application in multi-view settings often treats each view independently. The challenge of

how to effectively bridge and align the distinct, graph-informed latent spaces produced by multiple GVAEs remains a key area of research. Our framework uses GVAEs as the backbone for each view specifically to harness this relational information and then focuses on solving the subsequent alignment problem.

2.3. Contrastive Learning for Representation Alignment

Contrastive learning has emerged as a dominant paradigm for learning powerful self-supervised representations by maximizing the agreement between different augmentations or views of the same data instance [19, 30, 31, 32]. This principle has been extended to cross-modal settings to align representations from different data types, for example, between images and text [33, 34]. A critical finding in this domain is the utility of non-linear projection heads, which are applied after the primary encoder during training to create a space optimized for the contrastive loss. This decouples the representation learning for the contrastive task from that needed for downstream tasks, improving the quality of the final embeddings [19, 31, 35, 36]. However, the effective integration of projected contrastive learning with multi-view GVAE architectures, particularly in the presence of complex intra-modal structures like multiple lesions and significant data missingness, has not been fully explored. Our work, CLM-GVAE, pioneers this specific integration, using projected contrastive learning not as a pre-training step, but as a core, simultaneous objective to guide the GVAEs toward learning aligned, medically relevant representations.

3. Proposal Methods

We propose **CLM-GVAE**, an end-to-end framework designed to learn aligned, multi-modal patient representations for predicting immunotherapy response. Our approach addresses three key challenges: (1) capturing intra-modality relational structures and learning robust representations using view-specific Graph Variational Autoencoders (GVAEs); (2) handling intra-modality heterogeneity in radiology by introducing a hierarchical attention mechanism [37] for multi-lesion data; and (3) enforcing semantic coherence across modalities via a projected cross-view contrastive learning objective. The overall architecture is depicted in Figure 2. The model processes clinical, pathology, and radiology data through parallel streams before aligning their latent representations and fusing them for a final predictive task.

3.1. Dataset and Preprocessing

Our study utilizes a public multi-modal cohort of 366 NSCLC patients treated with PD-(L)1 blockade therapy from Vanguri et al. [20]. For our primary experiments and ablation studies, we focused on a well-defined TMB Sub-Cohort (N=247) containing patients with available genomic data. The Complete Cohort (N=366) was used to evaluate model robustness to extensive data missingness (see Figure 1 for modality distribution). The prediction task is binary immunotherapy response (Responder vs. Non-Responder) per RECIST v1.1.

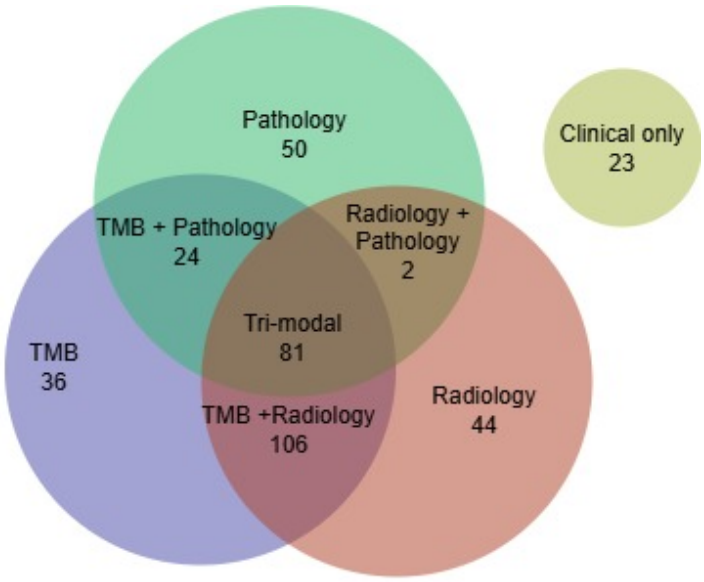


Figure 1: Distribution of Key Data Modalities Across the Patient Cohort.

3.1.1. Data Preprocessing and Feature Description

Clinical Data: Following median/mode imputation and scaling with RobustScaler, the clinical feature set included key demographic, laboratory, and treatment variables such as age, sex, ECOG performance status, smoking history (pack-years), baseline albumin, and dNLR.

Pathology Data: Patient-level pathology representations were derived from pre-extracted features from PD-L1 IHC stained slides. The final feature vector consisted of 137 features, including first-order statistics (e.g., mean, skewness) and texture features from Gray Level Co-occurrence Matrix (GLCM) analysis, such as Contrast, Correlation, Energy, and Homogeneity.

Radiology Data: To preserve intra-patient heterogeneity, we processed data at the lesion level. For each of the segmented lesions, a comprehensive set of 1,671 radiomics features was extracted using the pyradiomics package. These features included first-order statistics, shape-based features, and higher-order texture features from GLCM, GLRLM, GLSZM, NGTDM, and GLDM filters. These features were median-imputed and globally scaled.

Patient Similarity Graph Construction: To leverage relational information, we constructed a patient similarity graph for each view (clinical, pathology, radiology). For each view, we computed the cosine similarity between patient feature vectors and pruned the edges with similarity scores below a threshold of 0.7. For the radiology graph, patient-level similarity was computed using an aggregated signature of each patient's lesion statistics (mean, min, max, std). The resulting graphs serve as direct input to the GVAE encoders.

3.2. CLM-GVAE Model Architecture

The architecture of the CLM-GVAE is designed to process, align, and fuse multi-modal data in a principled manner. As illustrated in the high-level overview (Figure 2), the framework consists of three parallel streams for each data modality, a specialized module for handling multi-lesion radiology data, and a final stage for representation alignment and prediction. The following sections detail each of these key components.

3.2.1. Aggregation of Multi-Lesion Radiology

As depicted in Figure 3, to address the challenge of inter-lesion heterogeneity, we introduce a learnable attention module, the Attention-based Lesion Aggregator, to produce a single patient-level representation from a set of K_p individual lesion feature vectors $\{\mathbf{l}_{p,1}, \ldots, \mathbf{l}_{p,K_p}\}$. An attention score $e_{p,k}$ for each lesion is computed via a multi-layer perceptron (MLP), and these scores are normalized across all of the patient's lesions using a softmax function to obtain attention weights $\alpha_{p,k}$ by Eq 1.

$$\alpha_{p,k} = \frac{\exp(\mathrm{MLP}_{\mathrm{attn}}(\mathbf{l}_{p,k}))}{\sum_{j=1}^{K_p} \exp(\mathrm{MLP}_{\mathrm{attn}}(\mathbf{l}_{p,j}))} \tag{1}$$

Figure 2: **High-Level Data Flow:** Clinical, pathology, and radiology data are processed through parallel, view-specific GVAE streams. The resulting latent representations are then fused for prediction

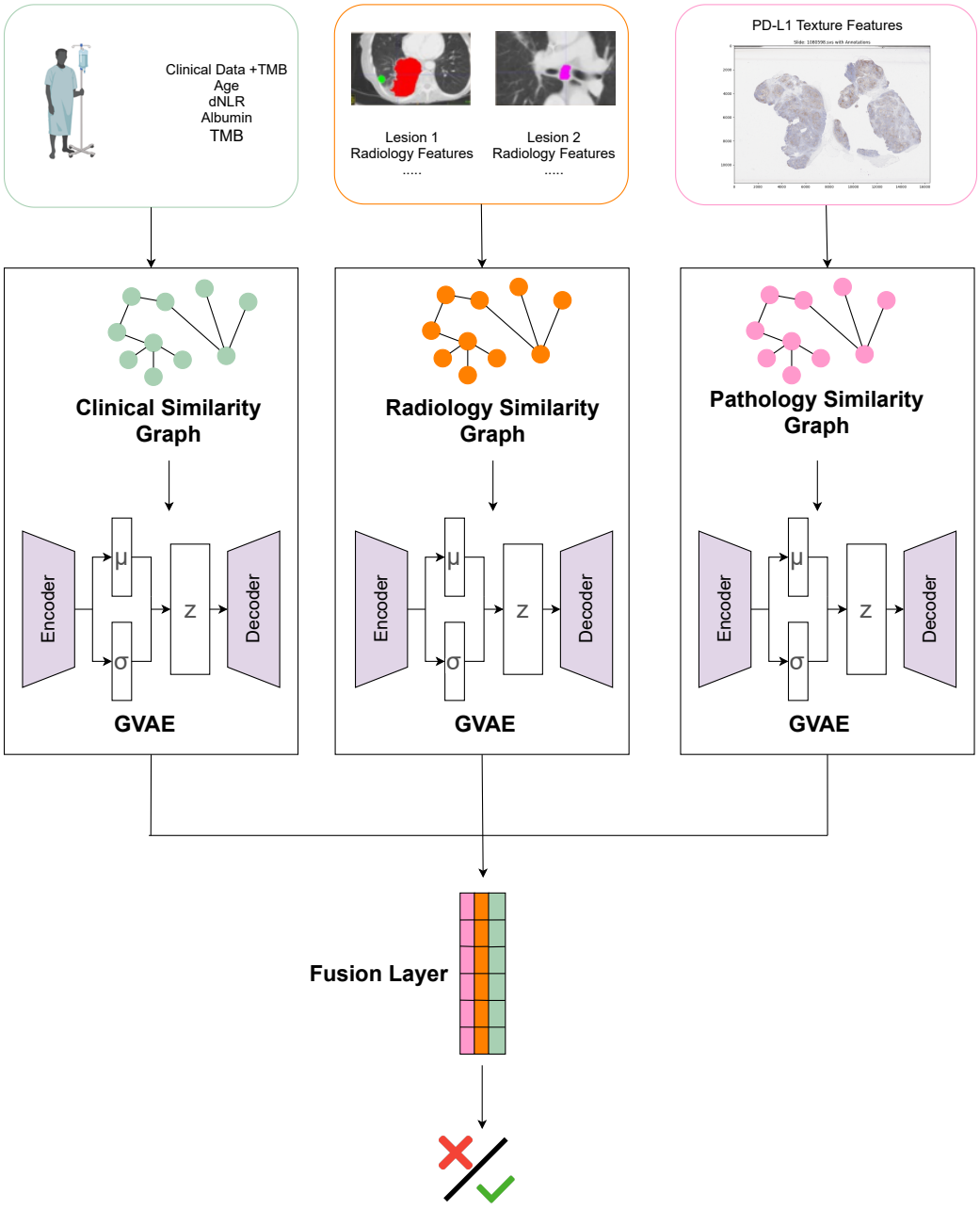
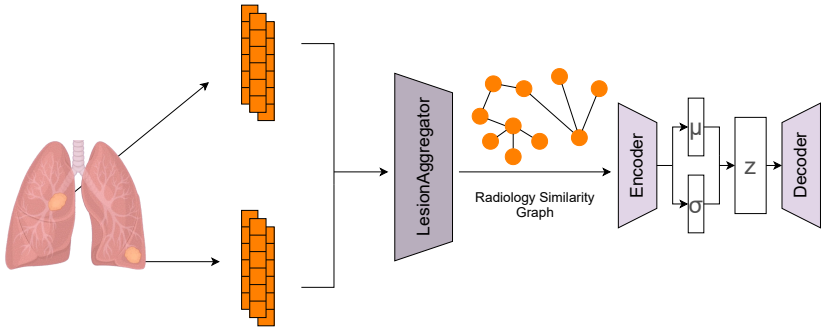


Figure 3: **Attention-based Lesion Aggregation:** A dedicated module for the radiology view uses an attention mechanism to aggregate features from multiple lesions into a single patient-level vector.



The final aggregated radiology feature vector for patient p, $\mathbf{x}_{p,\text{rad}}$, is the weighted sum of their lesion features, which then serves as the input to the radiology-specific GVAE by Eq 2.

$$\mathbf{x}_{p,\text{rad}} = \sum_{k=1}^{K_p} \alpha_{p,k}\, \mathbf{l}_{p,k} \tag{2}$$

3.2.2. View-Specific Graph Variational Autoencoders

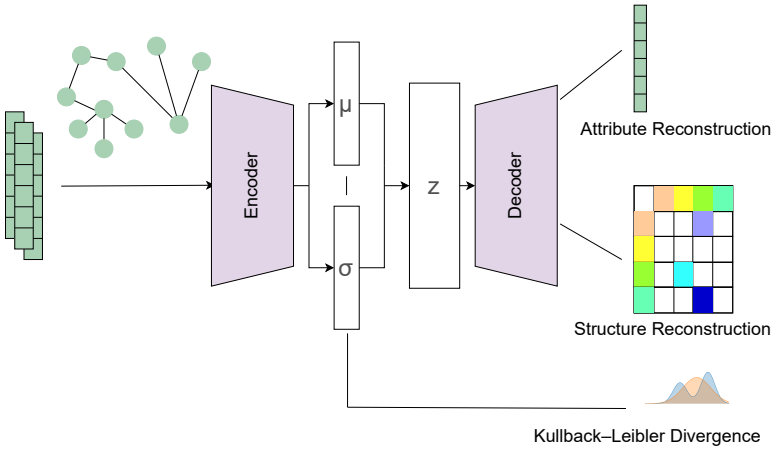
Each modality $v \in V$ (clinical, pathology, and aggregated radiology) is encoded by a dedicated GVAE, the core architecture of which is shown in Figure 4 .

- **Encoder:** The encoder is a multi-layer Graph Attention Network v2 (GATv2Conv) [17] that processes the patient features X_v and the corresponding similarity graph G_v. Its output is parameterized as a diagonal Gaussian distribution for each patient p, yielding a latent mean vector $\boldsymbol{\mu}_{v,p}$ and a log-variance vector $\log(\boldsymbol{\sigma}^2_{v,p})$. The approximate posterior is thus defined as by Eq 3.

$$q_{\phi_v}(\mathbf{z}_{v,p} \mid X_v, G_v) = \mathcal{N}(\mathbf{z}_{v,p} \mid \boldsymbol{\mu}_{v,p}, \text{diag}(\boldsymbol{\sigma}^2_{v,p})) \tag{3}$$

- **Decoder and VAE Objective:** For generative training, a latent vector $\mathbf{z}^{\sim}_{v,p}$ is sampled using the reparameterization trick [16]. As shown in Figure 4, a decoder network then aims to reconstruct both the input features $\mathbf{x}_{v,p}$ (via an MLP, measured by MSE loss, $\mathcal{L}_{\text{rec\_attr}}$) and

Figure 4: **Core GVAE Module:** For each view, a GNN encoder learns a probabilistic latent space (μ, σ), and a decoder reconstructs the original features and graph structure, regularized by a KL divergence loss.



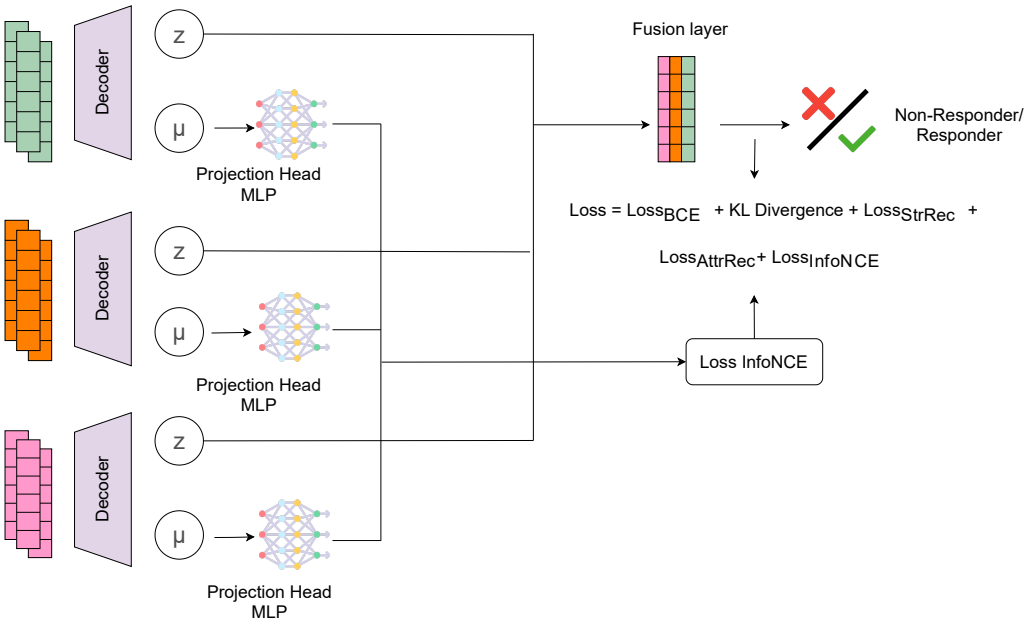
the graph structure A_v (via inner product of latent vectors, measured by BCE loss, $\mathcal{L}_{\text{rec\_struct}}$). The VAE objective is regularized by the KL divergence ($\mathcal{L}_{\text{kl}}$) between the approximate posterior and a standard normal prior.

3.2.3. Projected Cross-View Contrastive Alignment

To enforce semantic alignment between modalities, we introduce a cross-view contrastive learning objective. Following best practices in self-supervised learning [19, 36], detailed in the lower portion of Figure 5, we apply this loss in a separate projection space to avoid interfering with the VAE's reconstructive task.

- **Projection Heads:** The deterministic mean vector $\boldsymbol{\mu}_{v,p}$ from each GVAE is passed through a view-specific two-layer MLP projection head. The output is L2-normalized to produce the projected embedding $\boldsymbol{\mu}_{v,p}^{\text{proj}}$ used for contrastive loss calculation.
- **Contrastive Loss:** We treat representations from different views of the same patient as positive pairs, and all other samples in the batch as negatives. The InfoNCE loss [30] is used to maximize the similarity

Figure 5: **Alignment and Prediction:** The latent mean vectors (μ) are aligned using a projected cross-view contrastive loss (InfoNCE). The sampled latent vectors ($\mathbf{z}$) are integrated via a fusion layer to produce the final binary classification of patient response.



of positive pairs relative to negative pairs by Eq 4.

$$\mathcal{L}_{\text{cross-cl}} = -\mathbb{E}_{\mathcal{P}}\left[\log \frac{\exp(\text{sim}(\boldsymbol{\mu}_i^{\text{proj}}, \boldsymbol{\mu}_j^{\text{proj}})/\tau)}{\exp(\text{sim}(\boldsymbol{\mu}_i^{\text{proj}}, \boldsymbol{\mu}_j^{\text{proj}})/\tau) + \sum_{k\in\mathcal{N}_{i,j}} \exp(\text{sim}(\boldsymbol{\mu}_i^{\text{proj}}, \boldsymbol{\mu}_k^{\text{proj}})/\tau)}\right] \tag{4}$$

where $(\boldsymbol{\mu}_i^{\text{proj}}, \boldsymbol{\mu}_j^{\text{proj}})$ is a positive pair, $\mathcal{N}_{i,j}$ is the set of negatives, `sim` denotes cosine similarity, and τ is a temperature hyperparameter.

3.2.4. Multi-Modal Fusion and Prediction

The final steps of the framework, integration and prediction (Figure 5), are designed to synthesize the aligned representations into a final clinical outcome.

- **Missing Data:** To handle missing modalities, we substitute the absent view's latent vector with a dedicated, learnable embedding vector $\mathbf{e}_{\text{missing},v}$ which is optimized end-to-end.
- **Fusion Layer:** A complete set of embeddings per patient (either $\mathbf{z}_{v,p}^{\sim}$ from the VAE or $\mathbf{e}_{\text{missing},v}$) is treated as a sequence. A Transformer-based fusion module with a learnable `[CLS]` token [38] and multi-head self-attention aggregates this sequence into a single fused representation, $\mathbf{Z}_{\text{fused}}$.
- **Prediction Head:** A final MLP classifier takes $\mathbf{Z}_{\text{fused}}$ as input and outputs a single logit for predicting treatment response.

3.3. Training Details and Loss Function

The `CLM-GVAE` framework is trained end-to-end by minimizing a composite loss function, which is a weighted sum of the task-specific loss and the self-supervised objectives from the VAE and contrastive learning modules:

$$\mathcal{L}_{\text{total}} = w_{\text{class}}\mathcal{L}_{\text{BCE}} + w_{\text{cl}}\mathcal{L}_{\text{cross-cl}} + \frac{1}{|V|}\sum_{v\in V}(w_{\text{attr}}\mathcal{L}_{\text{rec\_attr},v} + w_{\text{str}}\mathcal{L}_{\text{rec\_struct},v} + w_{\text{kl}}\mathcal{L}_{\text{kl},v}) \tag{5}$$

The model was trained using the AdamW optimizer [39] with an initial learning rate of 1×10^{-4}. To stabilize training, the weights for the KL divergence (w_{kl}) and the contrastive loss (w_{cl}) were annealed from a low value to their target values over the initial 50-100 epochs. All experiments were conducted using 10-fold cross-validation.

4. Experiments and Discussions

This chapter presents the comprehensive experimental evaluation of the proposed CLM-GVAE framework. We assess the model's performance on the binary task of predicting immunotherapy response in NSCLC patients, first by comparing its performance against a suite of classical machine learning and state-of-the-art deep learning baselines. Subsequently, we present a series of detailed ablation studies to systematically dissect the contribution of each key architectural component.

4.1. Experimental Setup

To evaluate CLM-GVAE, we compared it against several baselines: (1) standard machine learning models (Logistic Regression, SVM, Random Forest, XGBoost) trained on a concatenated feature vector where lesion-level radiology features were mean-pooled and missing modalities were mean-imputed; and (2) a re-implementation of the state-of-the-art attention-based deep learning model (DyAM) from the original data source paper [20].

Model performance was primarily evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC) [40]. We also report Accuracy, Precision, Recall, and F1-Score for a complete picture. All metrics were computed using a 10-fold cross-validation scheme with a fixed random seed to ensure robust and reproducible evaluation across all comparative models.

The CLM-GVAE model was implemented in PyTorch and PyTorch Geometric. Key hyperparameters, such as learning rate (AdamW optimizer, initial rate 1e-4), contrastive temperature ($\tau\_\text{cross} = 0.1$), and loss weights, were determined via preliminary experiments on validation folds. We employed an annealing schedule for the KL divergence and contrastive learning loss weights to ensure training stability. A detailed list of hyperparameters is provided in the Supplementary Material. All reported results are the mean $\pm$ standard deviation across the 10 folds.

4.2. CLM-GVAE Outperforms Baseline Models

The primary evaluation was conducted on the TMB sub-cohort (N=247) to allow for a direct comparison of models utilizing all data modalities, including genomics. As shown in Table 1, our proposed CLM-GVAE framework achieved a mean AUC of 0.806 $\pm$ 0.088. This result significantly outperforms all classical machine learning baselines, with the strongest baseline

(XGBoost) achieving a mean AUC of 0.680. Furthermore, CLM-GVAE surpasses the performance of the re-implemented DyAM deep learning model (AUC 0.740) [20], demonstrating the benefit of our proposed architecture that combines explicit representation alignment with hierarchical feature processing.

Table 1: Performance Comparison of **CLM-GVAE** against Baseline and Ablated Models.

Model / Method	Mean AUC	F1-Score	Accuracy	Precision	Recall
Classical Baselines					
Logistic Regression	0.530 ± 0.201	0.765 ± 0.097	0.666 ± 0.120	0.787 ± 0.117	0.750 ± 0.098
Random Forest (RF)	0.676 ± 0.153	0.836 ± 0.081	0.739 ± 0.114	0.774 ± 0.106	0.915 ± 0.076
XGBoost	0.680 ± 0.182	0.819 ± 0.078	0.719 ± 0.106	0.777 ± 0.100	0.873 ± 0.084
Support Vector Machine (SVM)	0.677 ± 0.129	0.844 ± 0.065	0.735 ± 0.097	0.747 ± 0.095	0.978 ± 0.039
Deep Learning Baselines & Bimodal					
clinical only	0.717 ± 0.120	0.850 ± 0.064	0.755 ± 0.094	0.773 ± 0.095	0.955 ± 0.072
clinical + pathology	0.772 ± 0.125	0.843 ± 103	0.759 ± 0.126	0.793 ± 0.114	0.909 ± 0.132
clinical + radiology	0.729 ± 0.139	0.821 ± 0.121	0.746 ± 0.127	0.816 ± 0.113	0.835 ± 0.151
DyAM [20]	0.740 ± 0.106	0.773 ± 0.115	0.692 ± 0.108	**0.852 ± 0.071**	0.713 ± 0.115
Ablation Studies					
MLP CLM-GVAE (w/o GNN)	0.736 ± 0.079	0.858 ± 0.055	0.765 ± 0.075	0.771 ± 0.080	0.972 ± 0.044
CLM-GVAE (w/o CL)	0.761 ± 0.135	0.853 ± 0.089	0.767 ± 0.125	0.785 ± 0.103	0.939 ± 0.089
Proposed Full Model					
CLM-GVAE (Full)	**0.806 ± 0.088**	**0.869 ± 0.054**	**0.782 ± 0.079**	0.783 ± 0.071	**0.982 ± 0.056**

Figure 6 illustrates the consistent performance of **CLM-GVAE** across the 10 validation folds, with the mean ROC curve achieving an aggregated AUC of 0.81, indicating strong discriminative capability.

4.3. Ablation Studies Confirm the Importance of Core Components

To dissect the framework and validate our architectural choices, we conducted a series of ablation studies, with key results summarized in Table 1.

- Impact of Contrastive Learning: To quantify the impact of our alignment strategy, we trained a variant of the model without the cross-view contrastive loss (`CLM-GVAE w/o CL`). This resulted in the most significant performance degradation among the architectural ablations, with the mean AUC dropping to 0.761. This result provides strong evidence that explicit latent space alignment is critical to the framework's success, enabling a more effective fusion of information.

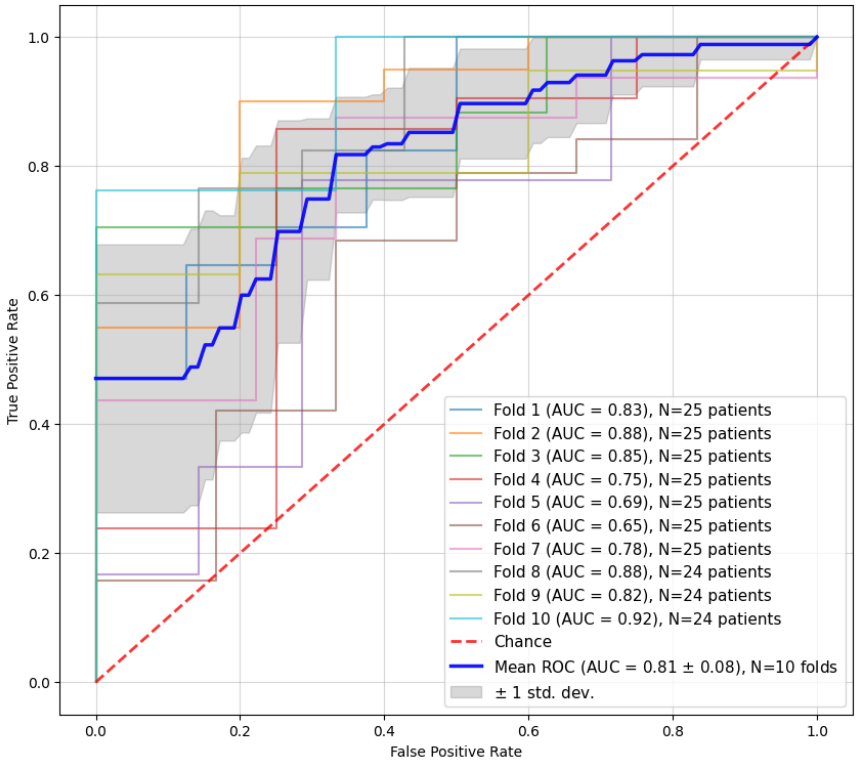


Figure 6: Receiver Operating Characteristic (ROC) Curve Analysis for `CLM-GVAE` on the TMB Sub-Cohort (N=247)

- Impact of Graph-based Encoding: We also assessed the value of modeling inter-patient relationships by replacing the GNN encoders with standard MLPs (`CLM-GVAE w/o GNN`). This model achieved a mean AUC of 0.736, markedly underperforming the full model. This demonstrates that contextualizing patient features with information from similar neighbors via GNNs produces richer latent representations that are more effective for both alignment and final prediction.

- Multi-modal Synergy: The results also show that bimodal combinations, while generally better than the `clinical only` baseline (AUC 0.717), could not match the performance of the full tri-modal integration. For instance, clinical + pathology reached an AUC of 0.772, but still fell short of the 0.806 achieved by CLM-GVAE. This suggests that

each modality provides unique and complementary predictive information that is effectively synergized by our fusion architecture.

4.4. Robustness to Extensive Missing Data

A crucial test of a clinical model's utility is its ability to handle real-world data sparsity. To evaluate this, we trained and tested the full CLM-GVAE framework on the complete cohort of 366 patients, which includes individuals with missing genomic and other data modalities. Leveraging its learnable missing embedding strategy, CLM-GVAE maintained robust performance, achieving a mean AUC of 0.711 $\pm$ 0.080. While this performance is, as expected, lower than on the more curated TMB sub-cohort, it remains a strong and clinically relevant result. This demonstrates the model's ability to gracefully handle data sparsity and generate predictions for all patients, a key feature for potential clinical translation. Full performance metrics and the corresponding ROC curve analysis for this cohort are provided in the Supplementary Material (Supplementary Table S1 and Supplementary Figure S1).

4.5. Hyperparameter Sensitivity Analysis

To justify the selection of key hyperparameters, we conducted a comprehensive series of sensitivity analyses. The detailed results for these studies are provided in the Supplementary Material. In summary, our investigation confirmed that the model's performance is robust across various architectural choices for the GVAE encoders, such as the embedding dimension and number of attention heads (Supplementary Table S2). The analysis of the contrastive learning objective identified a temperature of $\tau_{\text{cross}} = 0.2$ as optimal for achieving the highest discriminative performance (AUC), as shown in Supplementary Table S3. Furthermore, we found that using a stricter graph construction threshold of 0.8, which creates sparser graphs with high-confidence connections, significantly improved model performance by reducing noise during the GNN's message passing (Supplementary Table S4). These analyses collectively informed the final model configuration used for our primary experiments.

4.6. Discussions

The superior performance of CLM-GVAE can be attributed to several key architectural innovations. The most impactful of these, as confirmed by our ablation studies, is the projected cross-view contrastive learning objective.

The substantial drop in performance (from an AUC of 0.806 to 0.761) upon its removal validates the central hypothesis of this paper: explicitly enforcing semantic alignment between the latent spaces of different modalities is more effective than simply combining their features via attention or concatenation. This objective forces the model to learn a shared, synergistic representation that captures the underlying patient biology, rather than a weighted average of disparate signals. Furthermore, the demonstrated value of the GNN encoders (AUC dropping to 0.736 without them) indicates that patient-level context, derived from similarity to other patients, provides crucial relational information that is not present in a patient's features alone. The hierarchical lesion aggregator also likely contributes by producing a more nuanced radiological signature than simple feature pooling, a component whose individual contribution warrants further investigation.

Compared to the attention-based fusion approach of Vanguri et al. [20], which our model outperformed (AUC 0.806 vs. 0.740), our framework's explicit alignment offers a more principled approach to integration. While attention mechanisms effectively weigh the importance of modalities, our contrastive method ensures the representations are semantically compatible before fusion, leading to a more potent and informative fused embedding. Additionally, our model's strong performance on the complete $N = 366$ cohort (AUC 0.711) highlights the practical advantage of its learnable missing embedding strategy. This demonstrates a graceful degradation in performance as data becomes sparser, rather than a catastrophic failure, which is essential for models intended for real-world clinical application where complete data is a rarity.

5. Conclusions and Perspectives

In this study, we introduced CLM-GVAE, a novel multi-modal framework that integrates graph-based representation learning and contrastive alignment to predict immunotherapy response in NSCLC. Our results demonstrate that CLM-GVAE achieves state-of-the-art performance on this benchmark dataset, significantly surpassing both classical machine learning baselines and existing deep learning models. This work underscores the potential of building specialized, end-to-end architectures to unlock synergistic information from complex clinical data.

Despite the promising results, this study has several limitations. The model was developed and validated on a high-quality but single-institution

dataset; further external validation on multi-center cohorts is necessary to establish its generalizability and robustness to variations in data acquisition protocols. The observed variance in performance across cross-validation folds, while expected in clinical datasets of this size, suggests that future work could explore techniques to enhance model stability on more challenging data splits. Finally, while our lesion aggregator addresses intra-modality heterogeneity in radiology, extending this hierarchical approach to pathology (e.g., by processing whole-slide images at the patch level) could unlock further synergies.

In conclusion, CLM-GVAE presents a robust and effective framework for multi-modal data integration in oncology. By combining graph-contextualized generative representations with a powerful contrastive alignment strategy, our work provides a path toward building more accurate and reliable prognostic models. The principles demonstrated here-hierarchical feature processing, graph-informed encoding, and explicit cross-view alignment-offer a powerful blueprint for advancing precision medicine through the synergistic interpretation of complex patient data.

The framework and findings presented in this thesis open several exciting avenues for future research:

- External Validation and Model Harmonization: The most critical next step is to validate the CALM-VAE model on external, multi-institutional datasets. This would require addressing challenges related to data harmonization (e.g., different CT scanner protocols, different IHC staining procedures) and could involve techniques like federated learning to train models without sharing sensitive patient data.

- Advanced Fusion and Interpretability: Exploring more advanced fusion mechanisms, such as those that can explicitly model hierarchical relationships between modalities, could yield further performance gains. This could be coupled with the development of post-hoc explanation methods (e.g., Integrated Gradients, SHAP adapted for GNNs) to improve the clinical interpretability and trustworthiness of the model's predictions.

- Integration of Additional Modalities: The framework is inherently extensible. Future iterations could incorporate additional data types such as transcriptomics (RNA-seq), proteomics, or digital pathology features

beyond GLCM, potentially creating a more complete molecular and physiological picture of each patient.

- Longitudinal Data Analysis: The current model is based on baseline data. An important extension would be to adapt the architecture to handle longitudinal data (e.g., serial CT scans, changes in blood markers over time) using recurrent or temporal-attentive components to model disease progression and treatment response dynamically.

CRediT authorship contribution statement

Anh K. Le: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – Original Draft, Visualization. **Quang H. Nguyen**: Conceptualization, Methodology, Software, Resources, Writing - review & editing, Formal analysis, Validation, Visualization, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement

The code used in this study are available on our Github at `https://github.com/Kyanh56709/clm-gvae`.

References

[1] H. Mamdani, S. Matosevic, A. B. Khalid, G. Durm, S. I. Jalal, Immunotherapy in lung cancer: Current landscape and future directions, Front. Immunol. 13 (2022) 823618.

[2] A. Waqas, A. Tripathi, R. P. Ramachandran, P. A. Stewart, G. Rasool, Multimodal data integration for oncology in the era of deep neural networks: a review, Frontiers in Artificial Intelligence Volume 7 - 2024 (2024). doi:10.3389/frai.2024.1408843.
URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1408843

[3] K. M. Boehm, P. Khosravi, R. Vanguri, J. Gao, S. P. Shah, Harnessing multimodal data integration to advance precision oncology, Nature Reviews Cancer 22 (2) (2022) 114–126.

[4] P. Sharma, J. P. Allison, Immune checkpoint targeting in cancer therapy: toward combination strategies with curative potential, Cell 161 (2) (2015) 205–214.

[5] H. Yang, M. Yang, J. Chen, G. Yao, Q. Zou, L. Jia, Multimodal deep learning approaches for precision oncology: a comprehensive review, Briefings in Bioinformatics 26 (1) (2025) bbae699. arXiv:https://academic.oup.com/bib/article-pdf/26/1/bbae699/61330897/bbae699.pdf, doi:10.1093/bib/bbae699.
URL https://doi.org/10.1093/bib/bbae699

[6] S. Steyaert, M. Pizurica, D. Nagaraj, P. Khandelwal, T. Hernandez-Boussard, A. J. Gentles, O. Gevaert, Multimodal data fusion for cancer biomarker discovery with deep learning, Nature Machine Intelligence 5 (4) (2023) 351–362.

[7] H. Acosta, N. Henderson, J. Rowe, W. Min, J. Minogue, J. Lester, What's fair is fair: Detecting and mitigating encoded bias in multimodal models of museum visitor attention, in: Proceedings of the 2021 International Conference on Multimodal Interaction, ICMI '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 258–267. doi:10.1145/3462244.3479943.
URL https://doi.org/10.1145/3462244.3479943

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need (2023). arXiv:1706.03762.
URL https://arxiv.org/abs/1706.03762

[9] Y. An, H. Zhang, Y. Sheng, J. Wang, X. Chen, Main: Multimodal attention-based fusion networks for diagnosis prediction, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021, pp. 809–816. doi:10.1109/BIBM52615.2021.9669634.

[10] A. Gomaa, Y. Huang, A. Hagag, C. Schmitter, D. Höfler, T. Weissmann, K. Breininger, M. Schmidt, J. Stritzelberger, D. Delev, R. Coras,

A. Dörfler, O. Schnell, B. Frey, U. S. Gaipl, S. Semrau, C. Bert, P. Hau, R. Fietkau, F. Putz, Comprehensive multimodal deep learning survival prediction enabled by a transformer architecture: A multicenter study in glioblastoma, Neurooncol. Adv. 6 (1) (2024) vdae122.

[11] R. J. Chen, M. Y. Lu, W.-H. Weng, T. Y. Chen, D. F. Williamson, T. Manz, M. Shady, F. Mahmood, Multimodal co-attention transformer for survival prediction in gigapixel whole slide images, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3995–4005. doi:10.1109/ICCV48922.2021.00398.

[12] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: ICML, 2018, pp. 2132–2141.

[13] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, F. Mahmood, Data-efficient and weakly supervised computational pathology on whole-slide images, Nature Biomedical Engineering 5 (6) (2021) 555–570. doi:10.1038/s41551-020-00682-w.
URL https://doi.org/10.1038/s41551-020-00682-w

[14] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Transactions on Neural Networks 20 (1) (2009) 61–80. doi:10.1109/TNN.2008.2005605.

[15] T. N. Kipf, M. Welling, Variational graph auto-encoders (2016). arXiv:1611.07308.
URL https://arxiv.org/abs/1611.07308

[16] D. P. Kingma, M. Welling, Auto-encoding variational bayes (2022). arXiv:1312.6114.
URL https://arxiv.org/abs/1312.6114

[17] S. Brody, U. Alon, E. Yahav, How attentive are graph attention networks? (2022). arXiv:2105.14491.
URL https://arxiv.org/abs/2105.14491

[18] V. Petar, C. Guillem, C. Arantxa, R. Adriana, L. Pietro, B. Yoshua, Graph attention networks, in: International conference on learning representations, Vol. 8, 2018.

[19] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations (2020). arXiv:2002.05709.
URL https://arxiv.org/abs/2002.05709

[20] R. S. Vanguri, J. Luo, A. T. Aukerman, J. V. Egger, C. J. Fong, N. Horvat, A. Pagano, J. d. A. B. Araujo-Filho, L. Geneslaw, H. Rizvi, R. Sosa, K. M. Boehm, S.-R. Yang, F. M. Bodd, K. Ventura, T. J. Hollmann, M. S. Ginsberg, J. Gao, R. Vanguri, M. D. Hellmann, J. L. Sauter, S. P. Shah, MSK MIND Consortium, Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer, Nature Cancer 3 (10) (2022) 1151–1164.

[21] G. Muhammad, F. Alshehri, F. Karray, A. E. Saddik, M. Alsulaiman, T. H. Falk, A comprehensive survey on multimodal medical signals fusion for smart healthcare systems, Information Fusion 76 (2021) 355–375. doi:https://doi.org/10.1016/j.inffus.2021.06.007.
URL https://www.sciencedirect.com/science/article/pii/S1566253521001330

[22] S. G. Paul, A. Saha, M. Z. Hasan, S. R. H. Noori, A. Moustafa, A systematic review of graph neural network in healthcare-based applications: Recent advances, trends, and future directions, IEEE Access 12 (2024) 15145–15170. doi:10.1109/ACCESS.2024.3354809.

[23] X. Li, Y. Zhou, N. Dvornek, M. Zhang, S. Gao, J. Zhuang, D. Scheinost, L. H. Staib, P. Ventola, J. S. Duncan, BrainGNN: Interpretable brain graph neural network for fMRI analysis, Med. Image Anal. 74 (102233) (2021) 102233.

[24] J. Gao, T. Lyu, F. Xiong, J. Wang, W. Ke, Z. Li, Predicting the survival of cancer patients with multimodal graph neural network, IEEE/ACM Trans. Comput. Biol. Bioinform. 19 (2) (2022) 699–709.

[25] D. Ahmedt-Aristizabal, M. A. Armin, S. Denman, C. Fookes, L. Petersson, Graph-based deep learning for medical diagnosis and analysis: Past, present and future, Sensors 21 (14) (2021). doi:10.3390/s21144758.
URL https://www.mdpi.com/1424-8220/21/14/4758

[26] L. Liu, X. Zhu, J. Wu, Q. Hu, H. Cui, Z. Chen, T. Xu, Medical image processing using graph convolutional networks: A review, Perioper Precis Med 1 (2) (2023) 78–92.

[27] R. Wajgi, T. Champaneria, D. Wajgi, Y. Suryawanshi, D. Bhoyar, A. Nilawar, Heart disease prediction using graph neural network, International Journal of Intelligent Systems and Applications in Engineering 12 (12s) (2024) 280–287.
URL https://www.ijisae.org/index.php/IJISAE/article/view/4514

[28] H. Lu, S. Uddin, Disease prediction using graph machine learning based on electronic health data: A review of approaches and trends, Healthcare (Basel) 11 (7) (Apr. 2023).

[29] H. Oss Boll, A. Amirahmadi, M. M. Ghazani, W. O. de Morais, E. P. de Freitas, A. Soliman, F. Etminani, S. Byttner, M. Recamonde-Mendoza, Graph neural networks for clinical risk prediction based on electronic health records: A survey, Journal of Biomedical Informatics 151 (2024) 104616. doi:https://doi.org/10.1016/j.jbi.2024.104616.
URL https://www.sciencedirect.com/science/article/pii/S1532046424000340

[30] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding (2019). arXiv:1807.03748.
URL https://arxiv.org/abs/1807.03748

[31] T. Gao, X. Yao, D. Chen, SimCSE: Simple contrastive learning of sentence embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6894–6910. doi:10.18653/v1/2021.emnlp-main.552.
URL https://aclanthology.org/2021.emnlp-main.552/

[32] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2, 2006, pp. 1735–1742. doi:10.1109/CVPR.2006.100.

[33] R. J. Chen, M. Y. Lu, J. Wang, D. F. K. Williamson, S. J. Rodig, N. I. Lindeman, F. Mahmood, Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis, IEEE Trans. Med. Imaging 41 (4) (2022) 757–770.

[34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, Vol. 139 of Proceedings of Machine Learning Research, PMLR, 2021, pp. 8748–8763.
URL https://proceedings.mlr.press/v139/radford21a.html

[35] Z. Yang, H. Liu, X. Wang, A multimodal adaptive cooperative learning framework for cancer survival risk prediction, in: Proceedings of the 1st International Workshop on Multimedia Computing for Health and Medicine, MCHM'24, Association for Computing Machinery, New York, NY, USA, 2024, p. 64–69. doi:10.1145/3688868.3689195.
URL https://doi.org/10.1145/3688868.3689195

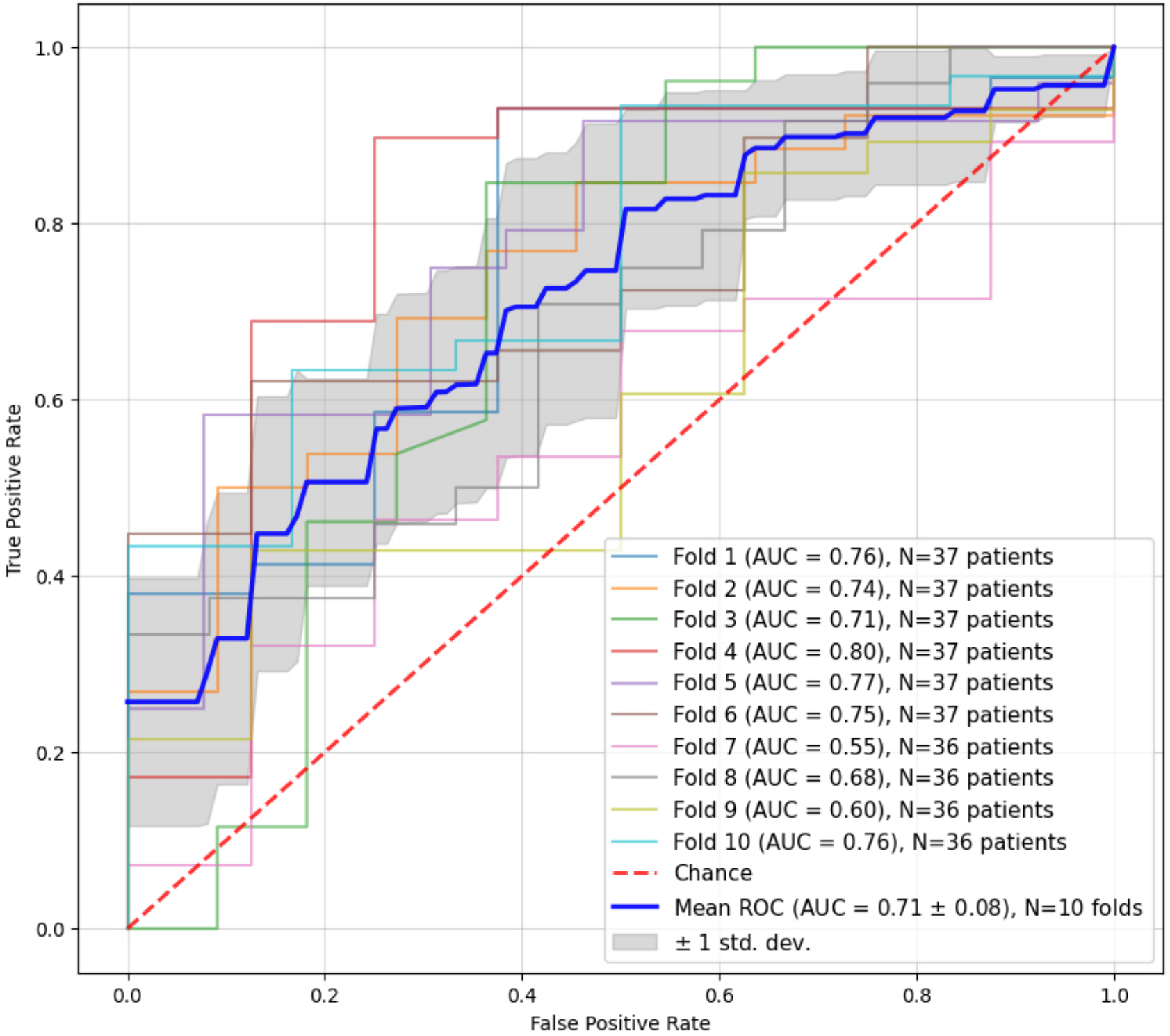
[36] Y. Gui, C. Ma, Y. Zhong, Unraveling projection heads in contrastive learning: Insights from expansion and shrinkage (2023). arXiv:2306.03335.
URL https://arxiv.org/abs/2306.03335

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.

[38] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding (2019). arXiv:1810.04805.
URL https://arxiv.org/abs/1810.04805

[39] I. Loshchilov, F. Hutter, Decoupled weight decay regularization (2019). arXiv:1711.05101.
URL https://arxiv.org/abs/1711.05101

[40] T. Fawcett, An introduction to roc analysis, Pattern Recognition Letters 27 (8) (2006) 861–874, rOC Analysis in Pattern Recognition. doi:https://doi.org/10.1016/j.patrec.2005.10.010.
URL https://www.sciencedirect.com/science/article/pii/S016786550500303X



Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Anh K. Le: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Visualization**. Quang H. Nguyen:** Conceptualization, Methodology, Software, Resources, Writing - review & editing, Formal analysis, Validation, Visualization, Supervision**.**