Mixture-of-Experts Guided Multi-Omic Integration for Gastrointestinal Cancer Subtype Prediction

Sajib Acharjee Dip[1]
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
sajibacharjeedip@vt.edu

Uddip Acharjee Shuvo[2]
Department of Computer Science and Engineering
University of Dhaka
Dhaka, Bangladesh

Dipanwita Mallick[3]
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh

Abrar Rahman Abir[4]
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Liqing Zhang[1,†]
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
lqzhang@vt.edu

Abstract

Accurate cancer subtype classification is a cornerstone of precision oncology, informing therapeutic decisions and improving prognostic assessment. Gastrointestinal adenocarcinoma (GIAC), however, presents a particularly challenging case due to its molecular heterogeneity and overlapping histological features. Traditional approaches based on single-omic biomarkers or naive multi-omic concatenation often fail to capture the complex interdependencies across genomic, epigenomic, and transcriptomic layers. We introduce **MoXGATE** (Mixture-of-Experts Guided Multi-Omic Integration), a deep learning framework that leverages modality-specific expert encoders, cross-attention fusion, and learnable modality weights to enable robust and interpretable integration of gene expression, DNA methylation, and miRNA profiles. By combining expert specialization with attention-driven fusion, MoXGATE effectively captures cross-omic dependencies while adaptively weighting each modality according to its predictive relevance. To address severe class imbalance in GIAC subtyping, we further incorporate focal loss, enhancing sensitivity to underrepresented subtypes. Comprehensive evaluation on TCGA GIAC demonstrates that MoXGATE achieves superior accuracy compared to state-of-the-art baselines, while ablation studies confirm the contributions of expert routing, cross-attention, and modality weighting. Moreover, transfer experiments on the TCGA BRCA cohort highlight the model's adaptability beyond GIAC, underscoring its generalizability to other cancer types.

[1]† Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) AI4Science

1 Introduction

Cancer subtyping is essential in precision oncology, as it informs targeted therapy decisions and improves patient outcomes [1, 19]. Gastrointestinal adenocarcinoma (GIAC), a heterogeneous class of malignancies, poses particular challenges for subtype classification due to its molecular complexity and overlapping clinical features [27, 23]. Conventional histopathology and single-omic biomarkers often fail to capture the breadth of tumor heterogeneity, highlighting the need for integrative, data-driven approaches [7, 11].

The advent of next-generation sequencing (NGS) has enabled large-scale multi-omic profiling, spanning gene expression (mRNA), DNA methylation, and miRNA signatures [30, 16, 3]. Multi-omic integration offers a more comprehensive view of tumor biology [2, 17], yet leveraging such data remains challenging. Existing methods struggle with modality heterogeneity, redundancy, and computational scalability, limiting their ability to extract robust cross-omic signals [28, 12, 31].

Recent deep learning approaches have advanced multi-omic cancer classification by employing attention or graph-based architectures. For example, moBRCA-net [4] employs self-attention with simple concatenation, which does not fully capture cross-modality dependencies. DeepMoIC [31] uses graph convolutional networks (GCNs) for pan-cancer analysis but is restricted to a small number of subtypes and incurs computational overhead from autoencoders. MOGONET [28] and MoGCN [12] apply graph-based fusion but depend heavily on well-defined similarity graphs, which are difficult to construct and sensitive to hyperparameters. Attention-based methods such as MMCA [29] have shown promise by modeling inter-modality alignment, yet most lack mechanisms to control modality imbalance or reduce redundancy.

In this work, we propose a **Mixture-of-Experts (MoE) attention framework** for multi-omic cancer subtyping, specifically focusing on GIAC. MoE architectures [9, 20, 6] enable expert specialization and selective routing, making them well-suited for heterogeneous biological modalities. Our model assigns each modality (gene expression, DNA methylation, miRNA) to a set of expert encoders, with a gating mechanism dynamically selecting expert contributions. The expert outputs are refined through self-attention [24], and modality-specific embeddings are integrated via cross-attention fusion [29], ensuring that cross-omic dependencies are explicitly modeled. To address class imbalance, which is a critical issue in cancer cohorts [25], we applied focal loss [13], improving the sensitivity of classification for minority subtypes.

Our contributions are as follows:

- **MoE-based multi-omic encoding:** We introduce modality-specific expert encoders with gating to promote specialization and diversity in feature extraction.
- **Attention-driven integration:** We refine expert outputs using self-attention and employ cross-attention to model interdependencies across modalities.
- **Adaptive fusion:** Learnable modality weights dynamically adjust the contribution of each omic source, reducing redundancy and highlighting complementary signals.
- **Robust classification under imbalance:** Focal loss improves classification performance for minority GIAC subtypes, mitigating skewed class distributions.
- **Generalizability:** Although designed for GIAC, the framework is adaptable to other cancers, demonstrating strong transferability across multi-omic datasets.

2 Methodology

2.1 Model Architecture

We propose a Mixture-of-Experts (MoE) based multi-omics fusion framework that combines expert-level specialization, self-attention encoding, cross-attention integration, and focal loss optimization. As illustrated in Figure 1, the framework is designed to efficiently encode high-dimensional omics data, capture interdependencies across modalities, and address class imbalance in cancer subtype prediction.

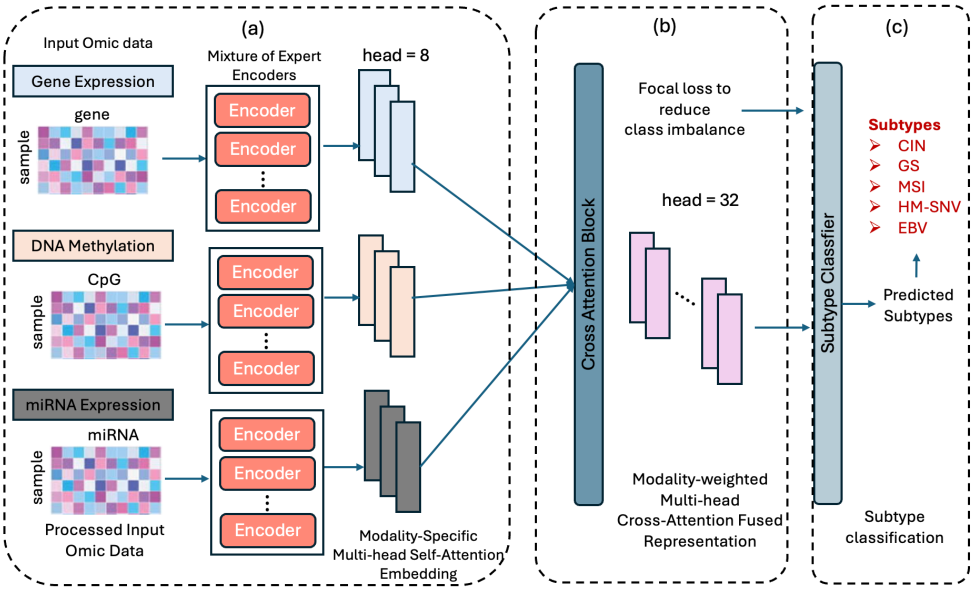

Figure 1: Overview of the proposed Mixture-of-Experts (MoE) based multi-omic cancer subtype classification framework. The architecture consists of three main components: (a) Modality-Specific Expert Encoders, where each modality (gene expression, DNA methylation, and miRNA expression) is processed by a set of expert networks. A gating function dynamically routes inputs to experts, and the aggregated expert outputs are further refined through modality-specific multi-head self-attention to capture diverse feature representations. (b) Cross-Attention Fusion, which integrates modality-specific embeddings via multi-head cross-attention to model interdependencies across omics layers. This mechanism allows the network to emphasize complementary signals and mitigate redundancy across modalities. (c) Subtype Classification Module, where the fused embedding is projected through a feed-forward classifier and optimized using focal loss [18], ensuring robustness under severe subtype imbalance.

2.1.1 Mixture of Expert Encoding for Modality-Specific Representations

For each modality $m \in \{1, \ldots, M\}$ and sample i, the input feature vector is $\mathbf{x}_i^{(m)} \in \mathbb{R}^{d_m}$. We define a set of L_m expert encoders $\{E_{m,j}\}_{j=1}^{L_m}$, each parameterized by $\Theta_{m,j}$. The latent representation is modeled as a mixture distribution:

$$p(\mathbf{h}_i^{(m)} \mid \mathbf{x}_i^{(m)}) = \sum_{j=1}^{L_m} g_{i,m,j}\, p(\mathbf{h}_i^{(m)} \mid \mathbf{x}_i^{(m)}, \Theta_{m,j}) \tag{1}$$

where gating weights $g_{i,m,j}$ are computed via a softmax over a gating network:

$$g_{i,m,j} = \frac{\exp(\mathbf{u}_{i,m}^\top \mathbf{v}_{m,j})}{\sum_{k=1}^{L_m} \exp(\mathbf{u}_{i,m}^\top \mathbf{v}_{m,k})} \tag{2}$$

with $\mathbf{u}_{i,m}$ being a learned query vector for sample i in modality m. The final aggregated representation is

$$\mathbf{H}_m = \left[\mathbf{h}_1^{(m)}, \ldots, \mathbf{h}_N^{(m)}\right]^\top \in \mathbb{R}^{N \times d}. \tag{3}$$

2.1.2 Self-Attention within Modalities

After expert aggregation, modality-specific representations are refined via self-attention:

$$\mathbf{Q}_m = \mathbf{H}_m \mathbf{W}_Q, \quad \mathbf{K}_m = \mathbf{H}_m \mathbf{W}_K, \quad \mathbf{V}_m = \mathbf{H}_m \mathbf{W}_V \tag{4}$$

$$\mathbf{A}_m = \text{softmax}\left(\frac{\mathbf{Q}_m \mathbf{K}_m^\top}{\sqrt{d}}\right) \tag{5}$$

$$\mathbf{Z}_m = \mathbf{A}_m \mathbf{V}_m + \mathbf{H}_m \tag{6}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$ are learnable parameters.

2.1.3 Cross-Attention Fusion

The modality-level outputs $\{\mathbf{Z}_m\}_{m=1}^{M}$ are concatenated:

$$\mathbf{C} = [\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_M] \in \mathbb{R}^{M \times N \times d} \tag{7}$$

Cross-attention is then applied:

$$\mathbf{Q}_c = \mathbf{C}\mathbf{W}_Q^c, \quad \mathbf{K}_c = \mathbf{C}\mathbf{W}_K^c, \quad \mathbf{V}_c = \mathbf{C}\mathbf{W}_V^c \tag{8}$$

$$\mathbf{A}_c = \text{softmax}\left(\frac{\mathbf{Q}_c\mathbf{K}_c^\top}{\sqrt{d}}\right) \tag{9}$$

$$\mathbf{F} = \mathbf{A}_c\mathbf{V}_c \tag{10}$$

2.1.4 Classification with Focal Loss

The fused representation is classified as:

$$\hat{\mathbf{y}} = \sigma\left(\mathbf{W}_f\mathbf{F} + \mathbf{b}_f\right) \tag{11}$$

where $\mathbf{W}_f \in \mathbb{R}^{d \times K}$ and K is the number of cancer subtypes.

We adopt Focal Loss to address class imbalance:

$$\mathcal{L}_{\text{focal}} = -\sum_{i=1}^{K} \alpha_i (1 - p_i)^\gamma y_i \log p_i \tag{12}$$

where p_i is the predicted probability for class i, y_i is the one-hot ground truth, α_i is a class-specific weight, and γ is the focusing parameter.

2.1.5 Overall Optimization and Training

The final loss combines focal loss with expert diversity regularization:

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda \sum_{m=1}^{M} \sum_{j \neq k} \left\| \mathbf{H}_{m,j}^\top \mathbf{H}_{m,k} \right\|_F^2 \tag{13}$$

where the second term encourages diversity among experts by penalizing highly correlated outputs.

2.1.6 Experimental Setup

Each modality (gene expression, DNA methylation, and miRNA expression) was modeled using a Mixture-of-Experts (MoE) design [20, 6], with multiple expert encoders assigned per modality. A soft gating mechanism determined how expert outputs were combined, ensuring that the model could adaptively emphasize different transformations of the same omic source. To further refine these modality-specific representations, we applied multi-head self-attention [24] within each modality, allowing the network to capture dependencies and reduce redundancy in high-dimensional input features.

The resulting modality embeddings were integrated through a cross-attention fusion layer [29], which enabled the model to capture interdependencies across heterogeneous omics views. This fused representation was then projected through a feed-forward classification network with non-linear activation and dropout regularization [21], before being mapped to the set of cancer subtype classes.

Optimization was carried out using AdamW with weight decay [15], together with focal loss [13] to mitigate class imbalance. In addition, we introduced a diversity regularization term on the experts [9, 26] to discourage them from collapsing to similar solutions and to promote complementary specialization.

A complete specification of hyperparameters, including the number of experts, attention heads, embedding dimensions, and training parameters, is provided in Table 1.

Table 1: Hyperparameters used in the Mixture-of-Experts based multi-omics classification framework.

Component	**Setting**
Number of experts per modality (L_m)	4 (for each of gene, methylation, miRNA)
Expert aggregation	Soft gating (sample-dependent weights)
Self-attention heads (per modality)	8
Dropout (intra-modality attention)	0.1
Cross-attention heads	32
Cross-attention embedding dimension	256
Final embedding dimension	128
Classifier activation	ReLU
Classifier dropout	0.3
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-2}
Loss function	Focal Loss ($\gamma = 2, \alpha = 1$)
Regularization	Expert diversity penalty ($\lambda = 0.01$)
Batch size	64
Training epochs	200

3 Results

3.1 Performance Comparison on GIAC Subtype Classification

Table 2 presents a comprehensive comparison between our proposed framework and a broad spectrum of baselines, ranging from classical machine learning classifiers to modern deep multi-omic integration models and ablated variants of our approach. Several important trends emerge.

First, classical baselines such as multilayer perceptrons (MLP) and support vector machines (SVM) achieved relatively poor performance (Accuracy 0.6757 / F1 0.5449), reflecting the difficulty of modeling high-dimensional, heterogeneous omic data without explicit mechanisms for feature interaction. Random Forest (RF) and simple ensemble strategies improved performance (Accuracy 0.8378 and 0.8270, respectively), but their gains remain limited by shallow feature integration. Deep matrix factorization (DMF) performed slightly better (Accuracy 0.7946 / F1 0.7884) but still lagged behind deep neural integration approaches.

Among deep multi-omic methods, the Multi-Modal Autoencoder (MAE) performed strongly (Accuracy 0.9081 / F1 0.8969), suggesting that reconstruction-based objectives can capture shared structure across modalities. However, the Multi-Modal Variational Autoencoder (MVAE) suffered from instability and weaker discriminative capacity (Accuracy 0.7622 / F1 0.7173). Graph-based integration with MOGONET also underperformed (Accuracy 0.6710 / F1 0.6710), likely due to sensitivity to graph construction and modality-specific noise. moBRCA-net, a recent attention-based model, achieved competitive performance (Accuracy 0.8837 / F1 0.8972) but still fell short of our approach.

Our ablation variants reveal the contributions of individual components. Simple concatenation of MoE outputs (Accuracy 0.8486 / F1 0.8518) was consistently weaker than cross-attention fusion (Accuracy 0.8739 / F1 0.8712), highlighting that explicitly modeling inter-modality dependencies is superior to naive feature aggregation. The removal of weighted focal loss degraded performance (Accuracy 0.8631 / F1 0.8676), confirming its role in handling severe class imbalance. Furthermore, replacing hard routing with soft routing in the MoE layer reduced accuracy (0.8739 vs. 0.9117), suggesting that selective expert allocation encourages complementary specialization among experts.

Finally, our full model, MoXGATE (Mixture-of-Experts Guided Attention), achieved the strongest overall results, with Accuracy 0.9117, F1 0.9104, Precision 0.9117, and Recall 0.9061. These improvements, while moderate compared to strong deep baselines such as MAE and self-attention + cross-attention, demonstrate the additive benefit of combining expert specialization, attention-driven fusion, and focal loss. Importantly, the gains are consistent across all metrics, indicating a balanced improvement in both sensitivity and precision. Nonetheless, the relatively narrow margins over strong baselines suggest that while MoXGATE is robust, further investigation is warranted into scalability, interpretability, and the trade-off between focal loss reweighting and potential overfitting to minority subtypes.

Table 2: **Model comparison with baselines and related methods.** Metrics are Accuracy, F1, Precision, and Recall. All models were trained and evaluated under the same experimental protocol.

Model	Accuracy	F1	Precision	Recall
Classical Machine Learning Baselines				
MLP	0.6757	0.5449	0.7809	0.6757
SVM	0.6757	0.5449	0.7809	0.6757
Random Forest (RF)	0.8378	0.8080	0.8474	0.8378
DMF	0.7946	0.7884	0.7859	0.7946
Ensemble	0.8270	0.7721	0.8555	0.8270
Deep Multi-Omic Integration Models				
MAE	0.9081	0.8969	0.9084	0.9081
MVAE	0.7622	0.7173	0.7138	0.7622
MOGONET	0.6710	0.6710	0.6240	0.6340
moBRCA-net	0.8837	0.8934	0.8892	0.8972
Ablation Variants				
MoE + Concat	0.8486	0.8486	0.8486	0.8518
Self-Attn + Cross-Attn	0.8955	0.8873	0.8955	0.8948
MoE (Soft routing) + Cross-Attn	0.8739	0.8612	0.8739	0.8712
MoE w/o weighted Focal Loss + Cross-Attn	0.8631	0.8541	0.8631	0.8676
MoXGATE (Ours)	**0.9117**	**0.9061**	**0.9117**	**0.9104**

Table 3: **Ablation on input modalities** (single run; Self-Attn + Cross-Attn encoders, MoE hard routing, weighted focal loss). Metrics are Accuracy, F1, Precision, and Recall.

Input Modality	Accuracy	F1	Precision	Recall
Gene	0.8757	0.8688	0.8757	0.8698
Methylation	0.8973	0.8806	0.8973	0.8881
miRNA	0.8162	0.7976	0.8162	0.7958
Gene + Methylation	0.8973	0.8805	0.8973	0.8884
Gene + miRNA	0.8703	0.8674	0.8703	0.8682
Methylation + miRNA	0.8919	0.8827	0.8919	0.8865
Gene + Methylation + miRNA	**0.9117**	**0.9104**	**0.9117**	**0.9061**

3.2 Impact of Single vs. Multi-Omic Modalities

The ablation study highlights several clear trends. To examine the contribution of each modality, we performed ablation experiments under a fixed training recipe: Self-Attention + Cross-Attention encoders, Mixture-of-Experts with hard routing, and weighted focal loss. The results (single run) are summarized in Table 3.

(1) Methylation remains the strongest individual signal. Among the single modalities, methylation achieved the best performance (Acc 0.8973 / F1 0.8806), slightly higher than gene expression and well above miRNA. This suggests that methylation carries the most subtype-discriminative information in our dataset. Gene expression also performed competitively, while miRNA consistently lagged, indicating it may contribute more as a complementary source rather than as a standalone predictor.

(2) Bimodal fusion does not always surpass unimodal baselines. Gene+Methylation did not improve over Methylation alone (Acc 0.8973 in both cases), and the F1 score even decreased marginally (0.8805 vs. 0.8806). Gene+miRNA also offered no clear advantage compared to gene expression by itself. By contrast, Methylation+miRNA showed a modest benefit (Acc 0.8919 / F1 0.8827), slightly outperforming the Methylation baseline in F1. These mixed outcomes suggest that combining modalities introduces interactions that are not uniformly beneficial, and that fusion mechanisms may require additional calibration to prevent strong modalities from being diluted by weaker ones.

(3) Trimodal fusion provides consistent gains. The full three-modality setup yielded the highest overall performance (Acc 0.9117 / F1 0.9104), clearly above any unimodal or bimodal setting. This supports the intuition that gene expression, methylation, and miRNA carry complementary information, and that cross-attention fusion can exploit this complementarity when all channels are present.

3.3 Performance on Other Cancer Data

Model	Accuracy	F1-Score
AE+Cross Attn	0.82	0.79
moBRCANet	0.87	0.86
Ours	**0.89**	**0.88**

Table 4: Performance comparison on breast cancer subtype classification. Our model achieves the best performance, demonstrating strong generalization.

To further validate the generalizability of our method, we conducted experiments on the TCGA-BRCA dataset shown in Table 4, which consists of 1,057 breast cancer samples. The dataset includes five intrinsic subtypes from the PAM50 classification: luminal A, luminal B, HER2 overexpression, basal-like, and normal-like cancers. We followed the same preprocessing steps as applied to the GIAC dataset, ensuring consistency across experiments. The dataset was split into 80% training and 20% testing, with 10% of the training data used for validation.

As shown in Table 3, our model achieves an accuracy of 0.89 and an F1-score of 0.88, outperforming existing approaches such as AE+Cross Attention (0.82 accuracy) and moBRCANet (0.87 accuracy). These results demonstrate that our modality-aware cross-attention approach effectively generalizes across different cancer types, reinforcing its robustness in multi-omic cancer subtype classification.

3.4 Effect of Expert Count on Model Performance

Table 6 reports the effect of varying the number of experts assigned to each modality. Several patterns are evident from this experiment.

When only a single expert is used, performance is already competitive, with an accuracy of 0.8955 and an F1 score of 0.8948. This indicates that the combination of self-attention within modalities and cross-attention across modalities captures a substantial amount of the available signal even without expert diversity.

The best results were obtained with two experts per modality, reaching an accuracy of 0.9117 and an F1 score of 0.9104. This suggests that a limited degree of expert specialization provides complementary representations that enhance the model's ability to distinguish cancer subtypes.

Increasing the number of experts beyond two did not yield further improvements. With four experts, performance declined slightly, and with eight experts the accuracy dropped to 0.8919. This decrease may reflect redundancy among experts, over-parameterization relative to the dataset size, or instability introduced by hard routing when too many experts compete for limited training data.

Overall, these results indicate that a moderate number of experts, specifically two per modality, offers the most effective balance between specialization and stability. Larger expert pools appear to introduce inefficiency without improving predictive power, underscoring the importance of tuning the expert count to the data scale and complexity of the classification task.

3.5 Impact of Reducing Feature Space Across Modalities

Table 10 in Appendix summarizes the effect of progressively reducing the number of features across the three modalities (DEG, DMR, DEM). The violin plots in Figure 2 illustrate the distribution of accuracy and F1 across five independent runs for each feature set size, while line plots with mean $\pm$ standard deviation are provided in the Appendix (Figures 5 and 6) as an alternative view of the same results.

The full feature sets (20,530 DEGs, 23,381 DMRs, and 746 DEMs) yield the strongest overall performance, with an accuracy of 0.9117 and an F1 score of 0.9061. As expected, reducing the feature space generally leads to a gradual decline in performance, consistent with the information loss incurred when fewer discriminative signals are available.

Intermediate feature sizes, however, sustain competitive results. For example, limiting each modality to 1,000 features still achieves accuracy of 0.8865 and F1 of 0.8824, only modestly lower than the

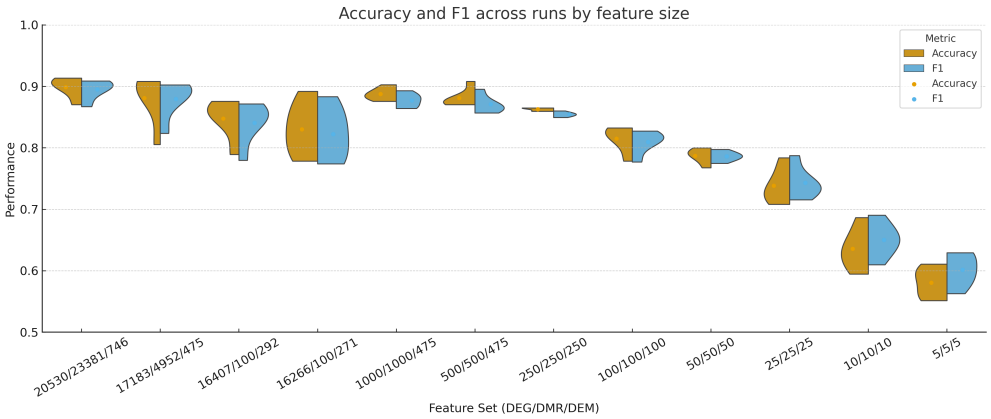

Figure 2: Performance variability across five runs for different feature set sizes. For each feature configuration (DEG/DMR/DEM), paired violin plots represent the distribution of Accuracy and F1-score. Violin width indicates the density of values, while overlaid dots mark run-level results. The plot highlights that full feature sets yield the strongest performance, but intermediate subsets (e.g., 500–1000 features per modality) preserve much of the predictive signal, whereas aggressive reductions lead to a systematic decline in both metrics.

Table 5: **Pathway overlap between predicted and real subtype enrichments.** Top-50 KEGG pathways (ranked by FDR) were compared, and Jaccard index was computed per subtype.

Subtype	#Predicted	#Real	#Overlap	%Overlap	Jaccard
EBV	50	50	50	100.0%	**1.0000**
GS	50	50	13	26.0%	0.1494
CIN	50	50	10	20.0%	0.1111
MSI	50	50	6	12.0%	0.0638
HM-SNV	50	50	4	8.0%	0.0417

full configuration. This suggests that a large fraction of the predictive power is concentrated in a smaller set of highly informative features. In contrast, very aggressive reductions (e.g., 50 features or fewer per modality) lead to sharp declines, with accuracy dropping below 0.77 and F1 below 0.78.

At mid-range feature sizes (e.g., 500–1,000 features), we also observe modest non-monotonic fluctuations. These are likely due to both training stochasticity and the heterogeneous informativeness of the selected features. The overall trend becomes more consistent at very small feature sets, where under-representation of critical biological signals is unavoidable.

Taken together, the results show that while maximal coverage of the omics space provides the strongest performance, much of the signal can still be retained with reduced feature sets. For practical applications where dimensionality is a concern, selecting on the order of 500–1,000 features per modality appears to strike a good balance between efficiency and predictive accuracy.

3.6 Pathway Enrichment and Biological Specificity Across Cancer Subtypes

To assess whether model-predicted subtypes recover biologically meaningful signatures, we performed KEGG pathway [10] enrichment analysis using GSEA (prerank) [22, 5] on the test-set gene expression profiles. Enrichment was computed separately for predicted and ground-truth subtype labels, and results were compared in terms of both overlap and pathway specificity.

3.6.1 Concordance Between Predicted and Real Subtypes

We first evaluated the overlap between predicted and real enrichment profiles by computing the Jaccard index [8] on the top-50 pathways per subtype (Table 5, Appendix Figure 3). The EBV subtype showed perfect concordance ($J = 1.0$), with all 50 predicted pathways overlapping those from the ground truth, reflecting the distinct viral-response signature characteristic of EBV-driven GIAC. CIN and GS subtypes demonstrated moderate agreement ($J = 0.11$ and $J = 0.15$, respectively), suggesting that the model partially captured their hallmark pathways but missed others. MSI and HM-SNV

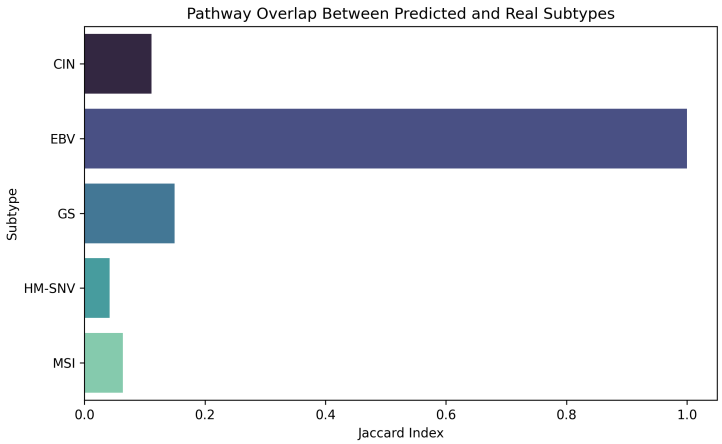

Figure 3: Pathway concordance between predicted and real subtype enrichments. The Jaccard index was computed between the top-50 KEGG pathways (ranked by FDR) for each subtype. EBV showed perfect overlap ($J = 1.0$), reflecting its distinct viral-response signature. CIN and GS exhibited moderate concordance ($J = 0.11$ and $J = 0.15$), while MSI and HM-SNV showed weak recovery ($J = 0.06$ and $J = 0.04$), consistent with their smaller sample sizes and greater intra-class variability.

showed weaker concordance ($J = 0.06$ and $J = 0.04$), which may be attributed to smaller sample sizes and higher intra-subtype heterogeneity. These results highlight that while MoXGATE preserves strong biological signal for distinct subtypes such as EBV, performance is uneven across rarer or more heterogeneous categories.

3.6.2 Pathway Specificity Across Subtypes

To identify which pathways most strongly discriminate between subtypes, we computed a Pathway Specificity Score (PSS), which quantifies the relative enrichment of a pathway in a single subtype compared to its variability across others. The top-20 pathways ranked by PSS are shown in Appendix Figure 7. Several biologically relevant categories emerge: immune processes (intestinal immune network for IgA production, hematopoietic cell lineage, proteasome), metabolic pathways (tyrosine metabolism, taurine and hypotaurine metabolism, drug metabolism), and genomic stability mechanisms (DNA replication). The enrichment of immune-related pathways is consistent with EBV tumors, which are characterized by strong antiviral and immune signatures. CIN and MSI subtypes, which are associated with chromosomal and microsatellite instability, showed enrichment in DNA replication and repair pathways, while GS subtypes were more strongly associated with metabolic processes such as glycosaminoglycan biosynthesis and choline metabolism.

Taken together, these analyses demonstrate that MoXGATE does more than label assignment: it preserves pathway-level signals that align with known biological hallmarks of GIAC subtypes. However, the low concordance for MSI and HM-SNV suggests that small sample sizes and high heterogeneity limit recovery of consistent pathways. Expanding sample cohorts, incorporating additional omics (e.g., proteomics), or applying subtype-aware regularization may further improve pathway-level fidelity.

4 Discussions

Our proposed Modality-Aware Cross-Attention model demonstrates state-of-the-art performance for multi-omic cancer subtype classification, effectively integrating heterogeneous omics data sources. The cross-attention mechanism, combined with learnable modality weights, enhances the fusion of gene expression, DNA methylation, and miRNA data, capturing intricate inter-modality dependencies. The ablation studies confirm that cross-attention outperforms simple concatenation, emphasizing its significance in multi-omic integration. Additionally, the results highlight the dominance of methylation and gene expression data in driving classification performance, aligning with biological insights into cancer heterogeneity. The strong generalization to breast cancer subtypes further underscores the robustness and transferability of our approach beyond gastrointestinal adenocarcinoma (GIAC).

Table 6: **Ablation on the number of MoE experts.** All models use self-attention within modalities, cross-attention for fusion, hard expert routing, and weighted focal loss. Metrics are Accuracy, F1, Precision, and Recall.

# Experts	Accuracy	F1	Precision	Recall
1	0.8955	0.8948	0.8955	0.8873
2	**0.9117**	**0.9104**	**0.9117**	**0.9061**
4	0.9009	0.8972	0.9009	0.8923
8	0.8919	0.8940	0.8919	0.8809

Despite these advancements, certain limitations persist. First, while cross-attention improves modality fusion, it inherently increases computational complexity, making it less scalable for ultra-large datasets. Additionally, although modality weights provide insight into the relative importance of omics data, they do not explicitly model dynamic feature importance at the patient level, potentially limiting interpretability for individualized cancer profiling. Future work should explore efficient self-attention mechanisms to reduce complexity and incorporate patient-specific attention weighting for improved personalization.

5 Acknowledgments

We thank Virginia Tech, the Department of Computer Science, and the Fralin Biomedical Research Institute at VTC for institutional resources and support. This work was supported by the U.S. National Science Foundation (NSF) under Awards #2125798, #2344169, and #2319522.

References

[1] Bo Ahren. Islet g protein-coupled receptors as potential targets for treatment of type 2 diabetes. *Nature reviews Drug discovery*, 8(5):369–385, 2009.

[2] Matteo Bersanelli et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics*, 17:S15, 2016.

[3] Kasit Chatsirisupachai, Tom Lesluyes, Luminita Paraoan, Peter Van Loo, and João Pedro De Magalhães. An integrative analysis of the age-associated multi-omic landscape across cancers. *Nature communications*, 12(1):2345, 2021.

[4] Joung Min Choi and Heejoon Chae. mobrca-net: a breast cancer subtype classification framework based on multi-omics attention neural networks. *BMC bioinformatics*, 24(1):169, 2023.

[5] Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 39(1):btac757, 2023.

[6] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[7] Claudio Isella, Francesco Brundu, Sara E Bellomo, Francesco Galimi, Eugenia Zanella, Roberta Porporato, Consalvo Petti, Alessandro Fiori, Francesca Orzan, Rebecca Senetta, et al. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. *Nature communications*, 8(1):15107, 2017.

[8] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

[9] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[10] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[11] Vessela N Kristensen, Ole Christian Lingjærde, Hege G Russnes, Hans Kristian M Vollan, Arnoldo Frigessi, and Anne-Lise Børresen-Dale. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5):299–313, 2014.

[12] Xiao Li, Jie Ma, Ling Leng, Mingfei Han, Mansheng Li, Fuchu He, and Yunping Zhu. Mogcn: a multi-omics integration method based on graph convolutional network for cancer subtype analysis. *Frontiers in Genetics*, 13:806842, 2022.

[13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[14] Yang Liu, Nilay S Sethi, Toshinori Hinoue, Barbara G Schneider, Andrew D Cherniack, Francisco Sanchez-Vega, Jose A Seoane, Farshad Farshidfar, Reanne Bowlby, Mirazul Islam, et al. Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer cell*, 33(4):721–735, 2018.

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[16] Joaquin Mateo, Lotte Steuten, Philippe Aftimos, Fabrice André, Mark Davies, Elena Garralda, Jan Geissler, Don Husereau, Iciar Martinez-Lopez, Nicola Normanno, et al. Delivering precision oncology to patients with cancer. *Nature medicine*, 28(4):658–665, 2022.

[17] Otília Menyhárt and Balázs Győrffy. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Computational and structural biotechnology journal*, 19:949–960, 2021.

[18] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017.

[19] Daniela Senft, Mark DM Leiserson, Eytan Ruppin, and Ze'ev A Ronai. Precision oncology: the road ahead. *Trends in molecular medicine*, 23(10):874–898, 2017.

[20] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[22] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[23] Carmine Valenza, Lorenzo Guidi, Elena Battaiotto, Dario Trapani, Andrea Sartore Bianchi, Salvatore Siena, and Giuseppe Curigliano. Targeting her2 heterogeneity in breast and gastrointestinal cancers. *Trends in Cancer*, 10(2):113–123, 2024.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[25] Christine Walko, Patrick J Kiel, and Jill Kolesar. Precision medicine in oncology: New practice models and roles for oncology pharmacists. *American Journal of Health-System Pharmacy*, 73(23):1935–1942, 2016.

[26] Kaixin Wang, Bingyi Kang, Jie Shao, and Jiashi Feng. Improving generalization in reinforcement learning with mixture regularization. *Advances in Neural Information Processing Systems*, 33:7968–7978, 2020.

[27] Qianqian Wang, Ganglei Liu, and Chunhong Hu. Molecular classification of gastric adenocarcinoma. *Gastroenterology research*, 12(6):275, 2019.

[28] Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature communications*, 12(1):3445, 2021.

[29] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950, 2020.

[30] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[31] Jiecheng Wu, Zhaoliang Chen, Shunxin Xiao, Genggeng Liu, Wenjie Wu, and Shiping Wang. Deepmoic: multi-omics data integration via deep graph convolutional networks for cancer subtype classification. *BMC genomics*, 25(1):1–13, 2024.

A Appendix

A.1 Dataset

A.2 GIAC Cancer and Subtypes

Gastrointestinal Adenocarcinomas (GIACs) include four major cancer types: Colon Adenocarcinoma (COAD), Rectum Adenocarcinoma (READ), Stomach Adenocarcinoma (STAD), and Esophageal Carcinoma (ESCA). These cancers exhibit distinct histopathological and molecular characteristics:

- **COAD (Colon Adenocarcinoma):** A common gastrointestinal malignancy characterized by chromosomal instability (CIN) and microsatellite instability (MSI), with additional classifications based on molecular features.
- **READ (Rectum Adenocarcinoma):** Similar to COAD but arises in the rectum, sharing molecular features but influenced by distinct anatomic and treatment considerations.
- **STAD (Stomach Adenocarcinoma):** A highly heterogeneous cancer associated with multiple subtypes, including Epstein-Barr virus (EBV)-associated tumors, MSI-high tumors, and genomically stable (GS) subtypes.
- **ESCA (Esophageal Carcinoma):** A rare but aggressive cancer exhibiting CIN and MSI features, often linked to environmental and genetic risk factors.

A.3 Dataset Statistics

The dataset used in this study is sourced from The Cancer Genome Atlas (TCGA) [30], containing multi-omic profiles for GIAC cancers. We specifically focus on molecular subtyping based on genetic and epigenetic alterations. The dataset includes the following samples:

Abbreviation	Study Name	Subtype Classification	Subtypes	Samples
COAD	Colon Adenocarcinoma	Molecular	CIN, GS, MSI, HM-SNV, EBV	341
ESCA	Esophageal Carcinoma	Molecular	CIN, GS, MSI, HM-SNV, EBV	79
READ	Rectum Adenocarcinoma	Molecular	CIN, GS, MSI, HM-SNV, EBV	118
STAD	Stomach Adenocarcinoma	Molecular	CIN, GS, MSI, HM-SNV, EBV	383

Table 7: GIAC Cancer Subtypes and Sample Distribution from TCGA. The four studied cancers include Colon Adenocarcinoma (COAD), Esophageal Carcinoma (ESCA), Rectum Adenocarcinoma (READ), and Stomach Adenocarcinoma (STAD), with five molecular subtypes.

A.4 Molecular Subtypes in GIACs

Molecular subtyping in GIACs has been extensively studied using gene expression, oncogenic pathways, and histopathological criteria. However, traditional clustering approaches often struggle with the biological complexity inherent to these cancers. Our study leverages genomic, epigenomic, and transcriptomic data to define robust molecular subtypes. [14]

Key Subtype Characteristics:

- **EBV+ (Epstein-Barr Virus Positive):** Predominantly found in stomach cancers, characterized by extensive DNA hypermethylation.
- **MSI (Microsatellite Instability):** Associated with defective DNA mismatch repair, leading to a high mutation burden.
- **HM-SNV(Hypermutated-Single Nucleotide Variants):** Defined by an SNV-predominant mutation profile, often linked to POLE mutations.
- **CIN (Chromosomal Instability):** Characterized by large-scale chromosomal alterations, frequently found in GIAC tumors.

- **GS (Genome Stable):** Lacks significant chromosomal aberrations, representing a smaller but distinct subset of tumors.

The dataset integrates multiple molecular modalities, including mutation profiles, copy-number variations, and DNA methylation, ensuring a comprehensive framework for subtype classification.

A.5 Subtype Distribution for the Combined GIAC Dataset

Table 8 reports the distribution of GIAC subtypes across training, validation, and test sets. For a visual summary, Figure 4 shows a grouped bar chart that makes the imbalance across subtypes explicit.

Table 8: Distribution of GIAC subtypes across training, validation, and test sets.

ID	Subtype	Train	Validation	Test	Total
0	CIN	450	50	125	625
1	EBV	22	2	6	30
2	GS	78	9	22	109
3	HM-SNV	13	2	4	19
4	MSI	99	11	28	138
	Total	662	74	185	921

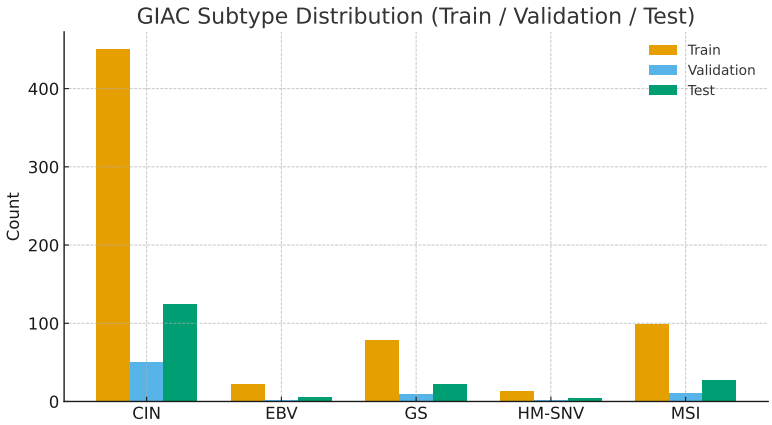

Figure 4: Grouped bar chart of GIAC subtype counts in the train, validation, and test splits. CIN is the dominant class, while EBV and HM-SNV are underrepresented.

B Data Processing Pipeline

To ensure a robust and unbiased evaluation, we utilized three cancer datasets (COAD, READ, STAD) for training and validation, while reserving the ESCA dataset exclusively for testing. Each cancer type in our dataset is categorized into five molecular subtypes. We performed a 90-10 split on the training dataset, where 90% of the samples were used for model training, and 10% for validation.

For feature preprocessing, we applied a two-step missing value handling strategy. First, we eliminated features with more than 40% missing values to ensure data reliability. Second, for the remaining missing values, we applied median imputation, filling in missing entries with the median value of the respective feature.

To maintain biological consistency across datasets, we selected only features that were common across all four cancer types. This yielded the following shared features:

- **Common Gene Expression Features:** 20,530
- **Common DNA Methylation Features:** 23,381
- **Common miRNA Features:** 746

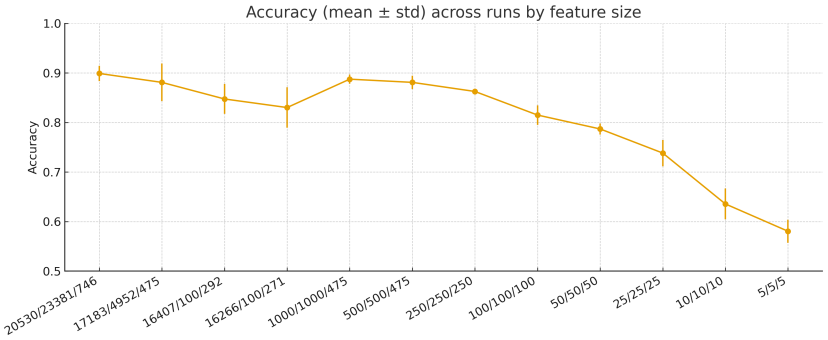

Figure 5: Accuracy (mean ± std) across five runs by feature size. Error bars denote standard deviation.

Following this preprocessing, our **final dataset** consisted of:

- **Training and validation set:** 842 samples
- **Test set (ESCA):** 79 samples
- **Final train-validation split:** 757 training samples and 85 validation samples

This data processing pipeline ensures that the model is trained on a diverse set of cancers while testing on a separate cancer type, providing a realistic evaluation of model generalizability across GIAC subtypes.

C Ablation Study

C.1 Ablation Study of Attention head

The number of heads in a multi-head cross-attention layer plays a critical role in capturing diverse feature interactions across omics modalities. The ablation study, as presented in Table 9, evaluates the performance of our model with 8, 16, and 32 attention heads. The results indicate that increasing the number of heads from 8 to 16 does not significantly impact performance, maintaining an accuracy of 94%. However, when the number of heads is increased to 32, the model achieves a slight improvement, reaching the highest accuracy of 95% along with a higher recall (0.95) and precision (0.96).

This improvement suggests that with a greater number of heads, the model is able to attend to finer-grained relationships among multi-omic features, thereby improving its ability to extract meaningful subtype-specific patterns. However, while a larger number of heads provides marginal gains, further increasing this number may introduce computational overhead without substantial performance benefits. Thus, 32 heads was selected as the optimal configuration, balancing both accuracy and computational efficiency.

Heads	Accuracy	Precision	Recall	F1-Score
8	0.94	0.96	0.94	0.94
16	0.94	0.96	0.94	0.94
32	0.95	0.96	0.95	0.94

Table 9: Ablation study on the effect of different numbers of heads in the cross-attention layer. The best-performing setting is highlighted.

C.2 Effect of Feature Set Size

Table 10 summarizes the effect of progressively reducing the number of features across the three modalities (DEG, DMR, DEM). The violin plots in Figure 2 illustrate the distribution of accuracy and F1 across five independent runs for each feature set size, while line plots with mean ± standard deviation are provided in the Appendix (Figures 5 and 6) as an alternative view of the same results.

Table 10: **Ablation on feature set size.** Performance reported as the average of 5 independent runs. Metrics are Accuracy, Precision, Recall, and F1-score. DEG = differentially expressed genes, DMR = differentially methylated regions, DEM = differentially expressed miRNAs.

DEG	DMR	DEM	Accuracy	Precision	Recall	F1
20530	23381	746	0.9117	0.9104	0.9117	0.9061
17183	4952	475	0.8973	0.8969	0.8973	0.8915
16407	100	292	0.8595	0.8595	0.8595	0.8543
16266	100	271	0.8324	0.8169	0.8324	0.8245
1000	1000	475	0.8865	0.8841	0.8865	0.8824
500	500	475	0.8757	0.8616	0.8757	0.8641
250	250	250	0.8649	0.8706	0.8649	0.8557
100	100	100	0.8324	0.8134	0.8324	0.8185
50	50	50	0.7676	0.7919	0.7676	0.7748
25	25	25	0.7514	0.7325	0.7514	0.7408
10	10	10	0.6324	0.6794	0.6324	0.6496
5	5	5	0.6108	0.6692	0.6108	0.6293

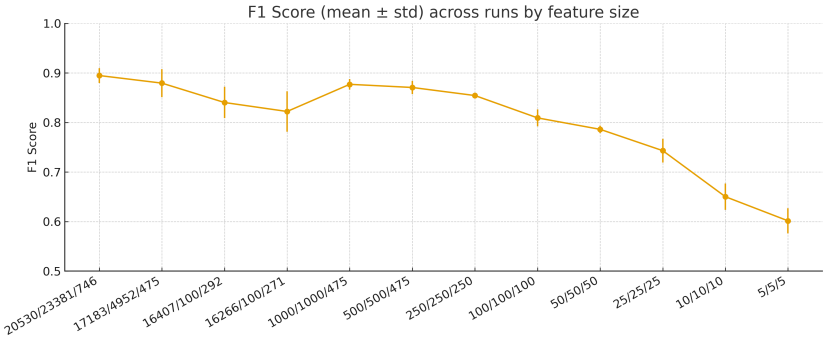

Figure 6: F1-score (mean ± std) across five runs by feature size. Error bars denote standard deviation.

C.3 Feature Selection Strategies

Table 11 summarizes the different criteria used to construct feature subsets for the ablation study. We varied statistical thresholds on differential expression and methylation (p-value, log fold change) to balance between sensitivity (capturing more candidate features) and specificity (focusing on fewer, high-confidence features). In addition, fixed top-k feature sets were evaluated for direct comparison.

To evaluate how feature dimensionality and statistical thresholds influence performance, we constructed multiple feature subsets using different selection criteria. For sensitivity-oriented strategies, we applied a lenient threshold (p-value ≤ 0.05, logFC ≥ 0.2), resulting in more than 17,000 genes and thousands of CpGs being retained. This maximizes coverage but may include weaker signals.

For balanced selection, we tightened the log fold change requirement (logFC ≥ 0.5), reducing CpGs to 100 and miRNAs to 292 while still retaining over 16,000 genes. Stricter specificity thresholds (p-value ≤ 0.01 or ≤ 0.001, logFC ≥ 0.5–1.0) further reduced the feature pool, trading breadth for high-confidence markers. Notably, the strictest criteria (p-value ≤ 0.001, logFC ≥ 1.0) eliminated all CpGs and miRNAs, leaving only 15,000 genes.

In parallel, we evaluated fixed top-k subsets (100, 250, 500, 1000 features per modality). These allow controlled comparisons where each modality contributes an equal number of features, providing insight into the trade-off between dataset size, computational efficiency, and classification performance. Together, these strategies enabled us to explore how both statistical filtering and dimensionality constraints affect downstream subtype prediction.

Table 11: Feature selection strategies and resulting feature counts across genes, CpGs, and miRNAs.

Strategy	p-value	logFC	Genes	CpGs	miRNAs
High Sensitivity (More Genes)	0.05	0.2	17183	4952	475
Moderate Sensitivity (Balanced Selection)	0.05	0.5	16407	100	292
High Specificity (Fewer, High-Confidence)	0.01	0.5	16266	100	271
High Specificity (Very Few, Strict Thresholds)	0.001	1.0	15178	0	0
Top 1000 Features	0.05	0.2	1000	1000	475
Top 500 Features	0.05	0.2	500	500	475
Top 250 Features	0.05	0.2	250	250	250
Top 100 Features	0.05	0.2	100	100	100

C.4 Additional Details on Pathway Enrichment Analysis

C.4.1 Methodological Details

For each GIAC subtype, we performed subtype-versus-rest enrichment using `GSEApy`'s prerank implementation of Gene Set Enrichment Analysis (GSEA) [22, 5]. Gene-level statistics were derived from log fold changes:

$$\Delta_k(g) = \log_2 \left(\frac{\overline{X}_k(g) + \epsilon}{\overline{X}_{\neg k}(g) + \epsilon} \right), \tag{14}$$

where $\overline{X}_k(g)$ is the mean expression of gene g among samples of subtype k, $\overline{X}_{\neg k}(g)$ is the mean among all other samples, and $\epsilon = 10^{-5}$ prevents division by zero. Genes were ranked by $\Delta_k(g)$ for each subtype, and enrichment was assessed against the KEGG 2021 Human pathways [10]. We restricted gene sets to size $5 \leq |S| \leq 500$ and used 100 permutations per run. Both nominal p-values and FDR q-values were reported, with top-20 pathways saved per subtype.

To evaluate subtype specificity, we constructed a pathway–subtype NES (normalized enrichment score) matrix and computed a Pathway Specificity Score (PSS):

$$\text{PSS}(p) = \frac{\max_k \text{NES}_{p,k} - \text{median}_k(\text{NES}_{p,k})}{\text{sd}_k(\text{NES}_{p,k}) + \delta}, \quad \delta = 10^{-6}, \tag{15}$$

ranking pathways that are highly enriched in one subtype relative to their variability across others.

C.4.2 Concordance Between Predicted and Real Subtypes

To test whether the model recapitulates real biological signals, we compared the top-50 predicted versus real enriched pathways for each subtype. Concordance was quantified using the Jaccard index:

$$J(k) = \frac{|\mathcal{P}_k^{\text{pred}} \cap \mathcal{P}_k^{\text{real}}|}{|\mathcal{P}_k^{\text{pred}} \cup \mathcal{P}_k^{\text{real}}| + \delta}. \tag{16}$$

Results (Table 5) show perfect recovery for EBV ($J = 1.0$), moderate recovery for CIN and GS ($J = 0.11$, $J = 0.15$), and weaker agreement for MSI and HM-SNV. These patterns are consistent with sample size effects (EBV being the most distinct, MSI/HM-SNV being smallest cohorts) and the biological distinctiveness of viral-driven tumors compared to genomically unstable ones.

C.4.3 Insights from Pathway Specificity

The PSS ranking (Figure 7) highlights several biologically coherent trends:

- **Immune-related pathways** (intestinal immune network for IgA production, hematopoietic cell lineage, proteasome) were enriched, particularly aligning with EBV subtypes where viral antigens drive immune activation.
- **Metabolic pathways** (tyrosine metabolism, taurine/hypotaurine metabolism, drug metabolism) appeared across CIN and GS, reflecting tumor metabolic rewiring.

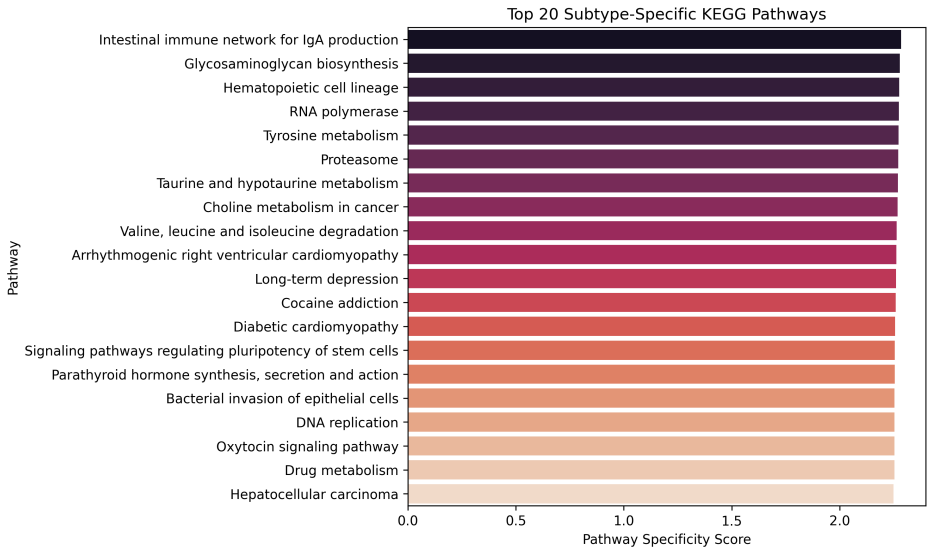

Figure 7: Top 20 subtype-specific KEGG pathways ranked by Pathway Specificity Score (PSS). PSS highlights pathways that show selective enrichment in one subtype compared to others. Immune-related and metabolic pathways (e.g., intestinal immune network for IgA production, proteasome, tyrosine metabolism) rank highly, reflecting biologically relevant processes underlying GIAC heterogeneity.

- **Genome stability pathways** (DNA replication, RNA polymerase, proteasome) were associated with MSI and CIN, consistent with their mutational instability profiles.
- **Cardiac and neuronal signaling pathways** (arrhythmogenic right ventricular cardiomyopathy, long-term depression, oxytocin signaling) also surfaced, which may reflect tissue-of-origin contamination or secondary effects rather than primary oncogenic mechanisms.

C.4.4 Limitations and Future Directions

While the enrichment results validate that MoXGATE predictions retain biologically coherent signals, several caveats remain:

1. **Permutation depth:** We used 100 permutations for efficiency, but more robust estimates would require 1,000+.
2. **Gene set choice:** KEGG [10] provides well-annotated metabolic and signaling pathways, but alternative databases (Reactome, Hallmark MSigDB) could yield complementary insights.
3. **Sample imbalance:** MSI and HM-SNV subtypes had fewer cases, which weakens pathway reproducibility. Future work could apply data augmentation or subtype-specific reweighting.
4. **Biological interpretation:** Some pathways with high scores may represent secondary effects or tissue background, not direct drivers of subtype biology. Careful curation is needed.

The pathway analysis confirms that MoXGATE does more than label assignment: it preserves subtype-distinguishing pathways, recapitulates known immune/viral hallmarks (EBV), and highlights instability-associated signatures (CIN, MSI). However, concordance is uneven across subtypes, and more systematic validation with larger, balanced cohorts and additional pathway collections will be essential for clinical translation.