GMR
The Original

HISTOMOLECULAR NET: A MULTI-SCALE DISENTANGLED AND CROSS-MODAL FRAMEWORK FOR PRECISION CANCER DIAGNOSIS

Manjula Jayamma[1], Mahankali Sankara Prasanna Kumar[2*]

[1]Research Scholar,Department of Computer Science and Engineering, Annamacharya University, New Boyanapalli, Rajampet, Annamayya District, Andhra Pradesh-516126, India, Email ID: jayamma.cse@srecnandyal.edu.in
[2*] Associate Professor, Department of Computer Science and Engineering, Annamacharya University, New Boyanapalli, Rajampet, Annamayya District, Andhra Pradesh-516126, India, Email ID:sankaraprasannakumar@gmail.com

ABSTRACT:
Cancers are highly heterogeneous and have varying prognostic outcomes, and clear classification is very essential in the patient stratification and to make clinical decisions. Although digital pathology has improved cancer diagnosis, modern day pathology seems to be no longer based on the sole use of histological characteristics but rather on the combination of molecular characterization. In order to meet this paradigm shift, we suggest a new framework of digital pathology that can simultaneously predict histology features and molecular markers and explicitly model the interaction between them. Multi-scale disentangling module is proposed to produce complementary representations with cellular level high-magnification and tissue level low-magnification whole slide images. Expanding on these characteristics, an attention-based hierarchical multi-task multi-instance learning framework is developed to do simultaneous histology and molecular prediction. Besides, it uses a co-occurrence probability-based label correlation graph and a cross-modal interaction module of dynamical confidence constraint and gradient modulation. The results of experiments show a high level of performance in comparison with state-of-the-art approaches, which indicates the potential for improving diagnostic accuracy in oncology of the framework.

KEYWORDS: Digital pathology, cancer classification, multi-scale learning, multi-task learning, multi-instance learning, molecular pathology, cross-modal learning, whole slide images, precision oncology

1. INTRODUCTION:

Cancer is a very heterogeneous disease; it has different histological phenotypes, molecular changes, and clinical outcomes. Correct classification of cancer is thus the key to patient stratification, estimation of prognosis and treatment plan. Historically, the diagnosis of cancer has been made by examination of the histopathological slides under the eyes of expert pathologists. The fact that the computational analysis of histology images is now possible on a large scale, due to the introduction of whole slide imaging (WSI) and digital pathology, has led to the creation of deep learning-based diagnostic systems [1], [2]. These approaches have already shown good results in tumor classification, grading, and outcome prediction of a variety of cancers [3], [6].

Notwithstanding these developments, modern cancer pathology has experienced a paradigm shift, which is driven by molecular profiling. The use of molecular markers, especially IDH mutation, 1p/19q co-deletion, and CDKN alterations have become more important in modern tumor classification, especially in glioma, and play an important role in both diagnosis and therapeutic decision-making [33]-[36]. Although the molecular testing is a valuable source of biological information, it is expensive, time-consuming and not always available. As a result, increasing attention is directed towards computational pathology procedures that would provide the ability to deduce molecular features using histology images only [11], [26], [27].

Recent work has examined deep learning-based models that predict molecular markers using WSIs that mostly rely on weakly supervised or multiple instance learning (MIL) models [5], [18], [22]. Nonetheless, the vast majority of current methods concentrate on either histology classification or on individual molecular markers alone, but do not explicitly reflect their intrinsic associations. In addition, WSIs have complicated hierarchical configurations of several magnifications, including cellular morphology to tissue properties. There is still a critical problem of integrating multi-scale information effectively because naive aggregation of features tends to result in the dilution of the information or dominance of the scale [8], [9].

Moreover, both histology and molecular features are naturally correlated, but their interactions are seldom represented in a principled fashion. Current strategies of multi-task learning tend to implicitly share representations, and do not impose biologically meaningful consistency or take advantage of co-occurrence patterns between molecular markers [20], [25]. This limitation limits the power of the existing systems to entirely encompass the multifaceted histology-molecular interaction that supports cancer heterogeneity.

To overcome these issues, we introduce a new digital pathology framework, which will simultaneously learn the multi-scale histology representations and molecular markers in a learning framework. The proposed solution fills the gap

between histological appearance and molecular characterization by extending existing studies by disentangling cross-magnification features, leveraging hierarchical attention based multi-task MIL, inclusion of label correlations and cross-modal interactions in the explicit model, and thus transforms precision oncology.

2. Related Work

Recent studies in digital pathology have studied deep learning techniques of analysing whole slide images to enable cancer diagnosis and prognostic analysis. Previously, there have been histology-based classification, multi-scale representation learning, multi-task modelling, and inferring molecular markers. Nevertheless, such initiatives are mostly disjointed and cross-scale context and histology-molecular interactions are rarely integrated, which drives the need to adopt integrated and biologically inspired learning models.

A. Histopathological Image Analysis via Deep Learning

Deep learning has radically transformed the analysis of histopathological images by removing human operators by automated identification of discriminative morphological patterns in digitized tissue slide [10]. The first methods were mostly patch-based and used convolutional neural networks in identifying local cellular and micro-architectural features of the presence of tumors, their subtype, and grade [8], [44]. Nevertheless, whole slide images (WSIs) need more than local patch reasoning, as clinically relevant knowledge is usually spread out and context-specific. To overcome the scarcity of dense annotations in WSI scale, weakly supervised learning emerged as the new paradigm, especially with multiple instance learning (MIL), in which a slide is represented as a bag of patches with slide-level supervision [6], [12]. Attention-based MIL goes further to enhancing this context by training instance importance weights to enable the model to pay attention to diagnostically salient regions, and provide post-hoc interpretability by attributing regions [5]. Although morphology-based tasks are well performed in cancers, the majority of WSI pipelines are trained to predict histology alone, and they do not directly encode molecular counterparts or histology-molecular interactions, which are gaining more and more prominence in clinical diagnostic processes [3], [11], [25].

B. Multi-Scale Representation Learning in Whole Slide Images

Whole slide images are hierarchically structured where diagnostically significant patterns are seen at various levels of spatial resolution, i.e. cellular morphology at high magnification and tissue-level structure at low magnification. To learn this complexity, multi-scale representation learning has been popularly studied in digital pathology [8], [9]. The present methods are usually to extract features in parallel at varying magnifications based on parallel convolutional branches or feature pyramid-based architectures, and then late fuse them to get slide-level representations [9], [18]. As much as these strategies raise the level of contextual awareness, they are usually based on basic concatenation or averaging functions and do not explicitly model scale dependencies. Consequently, one level of information can be overpowering or negating to the other level of information. Furthermore, most techniques implicitly assume scale consistency in a manner that is not taking into consideration scale-specific semantics or the cross-magnification flow of propagated information. The latter is especially an issue with WSIs, where cellular-level, as well as tissue-level, abnormalities are used together to diagnose them. As such, larger-scale learning processes must be more principled in that they need to be able to separate features that are scale specific and be able to interact across scale so that discriminative patterns across different resolutions can be used to form a coherent contribution to cancer classification [1], [2].

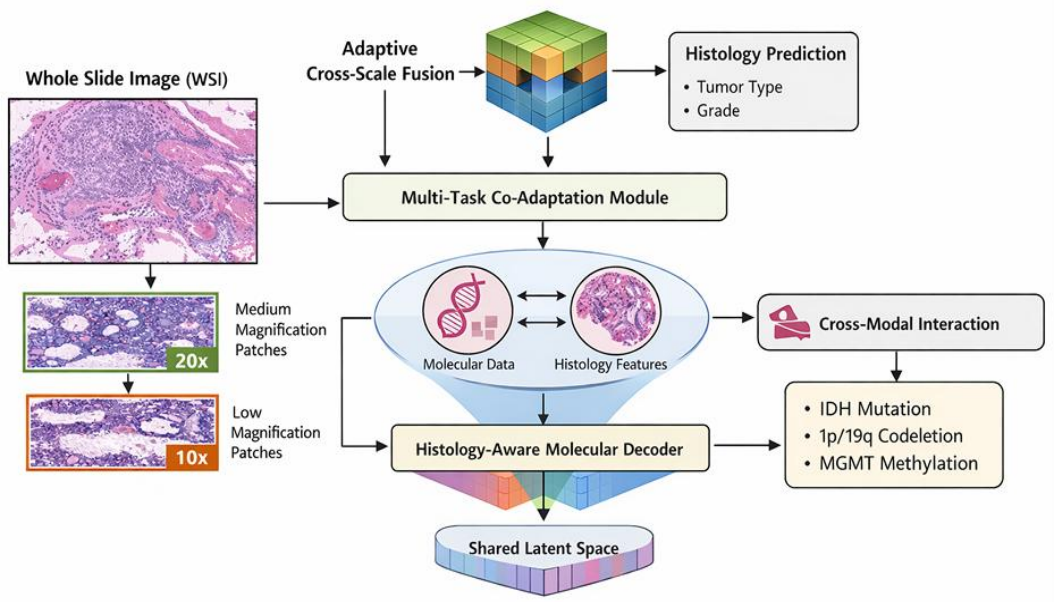

Fig. 1. Overall framework of the proposed multi-scale and cross-modal digital pathology model for cancer classification.

3. METHODOLOGY

A. Overall Framework

We introduce a unified multi-scale/cross-modal digital pathology platform based on joint histology classification and prediction of molecular markers with whole slide images. The framework works based on 10x and 20x representations, each with paired information on the tissue and cellular level. It contains four central components that include a multi-scale disentangling module, histology prediction module, molecular prediction module and cross-modal interaction module. The disentangling module is able to learn both shared and magnification specific features, which are scale aware, allowing effective propagation of cross-scale information. The prediction of histology is based on attention-based multiple instance learning, whereas molecular prediction is regarded as a structured multi-label task modelling the dependencies between the markers. Training is governed by cross-modal interaction mechanism that ensures consistency that is confidence-aware and that controls the interaction of tasks. The whole architecture is trained in an end-to-end fashion, which allows the integration of morphological and molecular patterns to effectively classify the cancers.

B. Multi-Scale Disentangling Module

Whole slide image (WSI) examination with a variety of magnifications is a regular part of clinical cancer diagnosis, with 10x magnification used to examine tissue-level architecture and 20x magnification used to detect cellular-level morphological information. Past biomedical studies denote that histology appearance and molecular markers correlate with unique but complementary multi-scale traits. In particular, molecular changes are frequently manifested by the fine-grained cell morphology, but histological phenotypes are more accurately described by the overall organisation of the tissues. Our observation is inspired by this and suggests a multi-scale disentangling module to explicitly factor out and model shared and task-specific representations across magnifications.

Let $\{X_i^{10}\}_{i=1}^{N}$ and $\{X_j^{20}\}_{j=1}^{N}$ denote patch sets extracted from a WSI at 10× and 20× magnifications, respectively. A pre-trained ResNet-50 backbone is employed to extract patch-level embedding:

$$\boldsymbol{F}^{10} \in \mathbb{R}^{N\times K}, \boldsymbol{F}^{20} \in \mathbb{R}^{N\times K},$$

Where K denotes the embedding dimension. In order to compensate the input of each magnification, the embeddings are scaled with learnable coefficients, α_{10} and α_{20}, and then concatenated:

$$\boldsymbol{F} = [\alpha_{10}\boldsymbol{F}^{10} \;\; || \;\; \alpha_{20}\boldsymbol{F}^{20}].$$

The fused representation, denoted as Fs is then fed through four separate multi-layer perceptron (MLPs) to produce shared and task representations of the fused representation of the molecular and histology prediction:

$$\boldsymbol{S}_m = \mathrm{MLP}_{S_m}(\boldsymbol{F}), \boldsymbol{I}_m = \mathrm{MLP}_{I_m}(\boldsymbol{F}),$$
$$\boldsymbol{S}_h = \mathrm{MLP}_{S_h}(\boldsymbol{F}), \boldsymbol{I}_h = \mathrm{MLP}_{I_h}(\boldsymbol{F}),$$

Where $\boldsymbol{S}_m$ and $\boldsymbol{S}_h$ denote shared representations, and $\boldsymbol{I}_m$ and $\boldsymbol{I}_h$ denote independent representations for molecular and histology tasks, respectively.

In order to achieve good disentanglement, we introduce cross-task disentanglement loss which reduces the difference between shared representations, and maximizes the distance between independent components:

$$\mathcal{L}_{\text{disent}} = \frac{\| \boldsymbol{S}_m - \boldsymbol{S}_h \|_2}{\| \boldsymbol{I}_m - \boldsymbol{I}_h \|_2 + \| \boldsymbol{I}_m - \boldsymbol{U}_m \|_2 + \| \boldsymbol{I}_h - \boldsymbol{U}_h \|_2}, \qquad (1)$$

Where $\boldsymbol{U}_m$ and $\boldsymbol{U}_h$ represent auxiliary latent anchors used to stabilize independent feature learning. This equation (1) promotes shared representations to represent scale-invariant semantics whereas independent representations hold task-specific and magnification-specific information. The resulting results of disentangled embedding give discriminative and task-relevant features to the following downstream histology and molecular prediction modules.

.

C. Molecular Prediction Module with Label Correlation Modelling

In modern time cancer diagnosis, molecular characterization is a key part especially in glioma classification where IDH mutation, 1p/19q co-deletion, and CDKN alterations are necessary markers to determine the subtype and prognosis. These molecular markers are not independent; rather they have high biological co-occurrence patterns. Such dependencies are thus important to model to make sound molecular inferences using histopathology images. Fig. 2 shows the detailed structure of the proposed molecular prediction module and its connection with the histology prediction branch.

Consideration of the disentangled multi-scale representations, obtained in Section III-B, the multi-label learning problem is developed to tackle the molecular prediction problem. Let

$$\boldsymbol{Z}_m \in \mathbb{R}^{N\times d}$$

Represent the molecular-oriented embedding generated by the multi-scale disentangling module, in which Nis the count of instances, and dis is the dimension of the embedding. These embeddings are sent to an organised molecular decoding pipeline, which is meant to model inter-marker relationships, as illustrated in Fig.2. In order to establish relationships between molecular markers, we determine a label

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}),$$

Where each node $v \in \mathcal{V}$ corresponds to a molecular marker, and the relations between statistically co-occurring molecular markers represented by edges.

The adjacency matrix

$$\boldsymbol{A} \in \mathbb{R}^{M \times M}$$

For Mmolecular markers is computed based on empirical co-occurrence probabilities:

$$A_{ij} = \frac{P(y_i = 1, y_j = 1)}{P(y_i = 1) + \epsilon},$$

Where $y_i$and $y_j$denote binary labels of the i-th and j-th molecular markers, respectively, and $\epsilon$is a small constant for numerical stability. The graph shown in Fig. 2 is the result of this formulation, the conditional dependency among the molecular markers.

Each molecular marker is associated with a learnable node embedding initialized from $\boldsymbol{Z}_m$. Graph convolution is then applied to propagate information across correlated markers:

$$\boldsymbol{H}^{(l|1)} = \sigma(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}),$$

Where $\widetilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$is the augmented adjacency matrix, $\widetilde{\boldsymbol{D}}$is the corresponding degree matrix, $\boldsymbol{W}^{(l)}$denotes learnable parameters, and $\sigma(\cdot)$is a nonlinear activation function. This graph-based propagation mechanism, depicted in **Fig. 2**, enables structured information exchange among molecular markers.

The refined node representations are subsequently passed through independent prediction heads to estimate molecular marker probabilities:

$$\widehat{\boldsymbol{y}}_m = \sigma(\boldsymbol{W}_m\boldsymbol{H}),$$

Where$\widehat{\boldsymbol{y}}_m \in [0,1]^M$. The molecular prediction loss is defined as the sum of binary cross-entropy losses across all markers:

$$\mathcal{L}_{\text{mol}} = \sum_{i=1}^{M} \text{BCE}(y_i, \hat{y}_i).$$

By explicitly modelling molecular label correlations and structurally integrating them within the molecular prediction pipeline illustrated in Fig. 2, this module enhances prediction consistency and biological plausibility, while providing structured molecular supervision for subsequent cross-modal interaction with histology prediction.

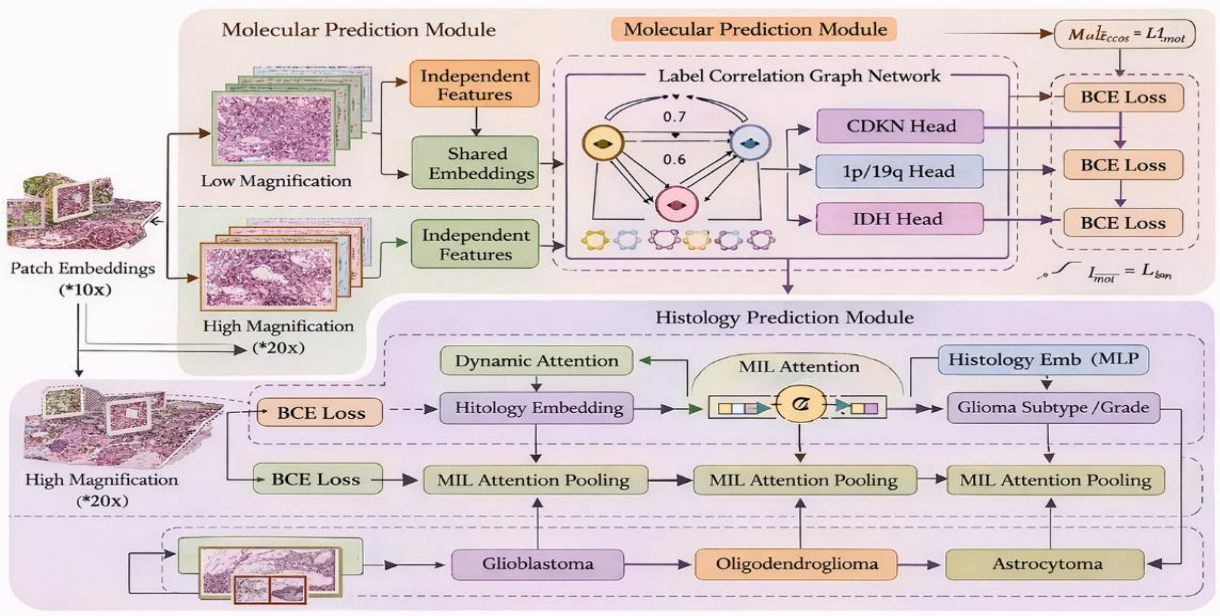

Fig.2 Detailed structure of the proposed molecular prediction module (above) and the histology prediction module (below).

D. Histology Prediction Module

Histological classification remains a fundamental component of cancer diagnosis, providing essential information on tumor subtype and grade. Unlike molecular prediction, which focuses on marker-specific dependencies, histology prediction requires effective aggregation of spatially distributed morphological patterns across a whole slide image (WSI). Given the weakly supervised nature of WSIs, where only slide-level labels are available, we formulate histology prediction as a **multiple instance learning (MIL)** problem.

Let

$$\boldsymbol{Z}_h \in \mathbb{R}^{N \times d}$$

denote the histology-oriented embeddings generated by the multi-scale disentangling module described in Section III-B, where $N$is the number of instances (patches) and $d$is the feature dimension. As illustrated in **Fig. 2**, these embeddings are processed by an attention-based MIL framework to produce a slide-level histology representation.

Specifically, an attention scoring function is applied to each instance embedding:

$$a_i = \frac{\exp(\boldsymbol{w}^\top \tanh(\boldsymbol{W}\boldsymbol{z}_{h,i}))}{\sum_{j=1}^{N} \exp(\boldsymbol{w}^\top \tanh(\boldsymbol{W}\boldsymbol{z}_{h,j}))},$$

where $\boldsymbol{z}_{h,i}$denotes the i-th instance embedding, $\boldsymbol{W}$and $\boldsymbol{w}$are learnable parameters, and $a_i$represents the normalized attention weight reflecting the contribution of the i-th patch to the slide-level decision.

Using the learned attention weights, the slide-level histology representation is computed as:

$$\boldsymbol{z}_{\text{his}} = \sum_{i=1}^{N} a_i \, \boldsymbol{z}_{h,i}.$$

This aggregation mechanism enables the model to focus on diagnostically salient regions while suppressing irrelevant or noisy patches, as visualized in **Fig. 2**.

The aggregated representation $\boldsymbol{z}_{\text{his}}$is subsequently passed through a multilayer perceptron (MLP) to predict the histology label:

$$\hat{\boldsymbol{y}}_{\text{his}} = \text{Softmax}(\boldsymbol{W}_{\text{his}}\boldsymbol{z}_{\text{his}}),$$

where $\hat{\boldsymbol{y}}_{\text{his}}$denotes the predicted probability distribution over histological classes. The histology prediction loss is defined using categorical cross-entropy:

$$\mathcal{L}_{\text{his}} = \text{CE}(\boldsymbol{y}_{\text{his}}, \hat{\boldsymbol{y}}_{\text{his}}).$$

Through attention-based MIL, the suggested histology-forecasting module successfully encodes heterogeneous morphologic patterns throughout the WSI. This design, combined with the molecular prediction module shown in Fig. 2, can be used to do coherent multi-task learning and to supervise further cross-modal interaction modeling.

E. Overall Objective Function and Optimization

In accordance with the multi-scale multi-task model that is proposed, the ultimate training goal will be to optimally predict histology, molecular markers and glioma classification, and explicitly capture cross-modal relationships between histology and molecular features. This formulation of the objective as a unified objective directly contributes to the performance increase that is later reported in Table I, with the proposed method showing superiority to competing methods in a variety of evaluation metrics in both datasets.

1) Task-Specific Loss Functions

The overall learning objective integrates multiple task-specific losses:

Histology prediction loss.

For histology feature prediction (i.e., necrosis and microvascular proliferation), a standard cross-entropy loss is applied:

$$\mathcal{L}_{his} = \text{CE}(\hat{y}_{his}, y_{his})$$

where $\hat{y}_{his}$and y_{his}denote the predicted and ground-truth histology labels, respectively.

Molecular marker prediction loss. The molecular prediction task is formulated as a multi-label classification problem for IDH mutation, 1p/19q co-deletion, and CDKN homozygous deletion. Binary cross-entropy loss is used for each marker:

Glioma classification loss

For the final glioma subtype classification, a categorical cross-entropy loss is employed:

$$\mathcal{L}_{glio} = \text{CE}(\hat{y}_{glio}, y_{glio})$$

The joint optimization of these task-specific losses enables the model to leverage complementary supervision signals, which is reflected in the superior accuracy, AUC, and F1-score values reported for glioma classification in **Table I**.

2) Disentanglement and Label-Correlation Regularization

The multi-scale disentanglement loss, which is defined in Section III-B, is added to the baseline to achieve effective separation and sharing of the features between tasks and improve the task-relevant representations and to reduce the redundancy. Also, the loss of label correlation, denoted as: $\mathcal{L}_{LC}$, is based upon the co-occurrence probability-based label-correlation graph and it is used to ensure consistency between molecular predictions by matching feature similarity to biologically relevant co-occurrence statistics. This regularization is part of the enhanced sensitivity and specificity of Table I relative to those methods, which do not explicitly model the label-correlation.

3) Constraints of cross-modal interaction

In order to explicitly model the relationship between histology and molecular predictions, the dynamic confidence constraint (DCC) loss, denoted as: $\mathcal{L}_{DCC}$ is presented (Section III-D). This loss informs the model to put emphasis on the spatial regions that are informative together between the modalities to support biologically meaningful associations like the co-occurrence of IDH wildtype and necrosis-related histology patterns.

Moreover, the cross-modal gradient modulated (CMG-Modu) approach is a dynamic calibration mechanism of the optimization procedure through which the prevailing modality, which is established by histology grading indicators, can guide changes in the gradient. This corresponds to less modality-specific bias and better diagnostic consistency which helps to explain the equal sensitivity and specificity enhancements in datasets found in Table I.

4) Final Optimization Objective

The entire training goal can be defined as a loss term weighted average of all terms of loss:

$$\mathcal{L}_{total} = \mathcal{L}_{glio} + \lambda_{his}\mathcal{L}_{his} + \lambda_{mol}\mathcal{L}_{mol} + \lambda_{dis}\mathcal{L}_{disent} + \lambda_{LC}\mathcal{L}_{LC} + \lambda_{DCC}\mathcal{L}_{DCC}$$

where λ_{his}, λ_{mol}, λ_{dis}, λ_{LC}, and $\lambda_{DCC}$are hyperparameters controlling the relative contribution of each component.

Such coordinated goal makes sure that histology, molecular markers and glioma classification are optimized in a coordinated and guaranteed way hence the consistent performance gains of all the evaluation metrics illustrated in Table I.

5) Training Strategy

The end-to-end training of network is performed by stochastic gradient descent with adaptive optimization. The CMG-Modu strategy is implemented dynamically during the training to control the gradient flow between histology and molecular branch to ensure compliance with the recent clinical diagnostic standards.
This integrated optimization model allows the model to learn simultaneously strong multi-scale representations, learn biologically significant label associations, and learn accurate glioma classification using whole slide images alone as quantitatively validated by the comparative outcomes found in Table I.

Table I. Mean values (in %) for glioma classification metrics on internal and external validation datasets

Approaches	**MTL***	**CMM**	**TCGA GBM-LGG + CPTAC**					**IvYGAP**				
			Acc.	Sen.	Spec.	AUC	F1-score	Acc.	Sen.	Spec.	AUC	F1-score
VGG			71.1	71.1	85.2	84.9	71.1	89.4	89.4	81.6	91.8	89.4
Inception			60.2	60.2	84.3	81.4	60.2	67.2	67.2	99.5	84.5	67.2
DenseNet			63.2	63.2	91.6	84.8	63.2	68.3	68.3	99.3	73.7	68.3
AlexNet			53.3	53.3	89.0	71.0	53.3	80.2	80.2	99.6	89.9	80.2
ResNet			65.5	65.5	52.7	80.4	65.5	55.3	55.3	63.5	94.6	55.3
ABMIL	✓		69.4	69.4	82.3	85.9	69.4	93.1	93.1	81.7	97.9	93.1
CLAM	✓		63.5	63.5	91.1	83.7	63.5	82.2	82.2	99.9	86.4	82.2
TransMIL	✓		66.1	66.1	84.0	85.6	66.1	82.3	82.3	69.3	93.1	82.3
Charm	✓		57.6	57.6	83.0	79.6	57.6	54.3	54.3	81.0	64.4	54.3
Deepglioma	✓		55.3	55.3	84.9	79.9	55.3	66.9	66.9	99.7	84.6	66.9
MCAT	✓	✓	68.0	68.0	89.3	87.7	68.0	–	–	–	–	–
CMTA	✓	✓	66.3	66.3	88.8	87.0	66.3	–	–	–	–	–
Wang et al.†	✓	✓	73.0	73.0	82.3	87.2	73.0	95.7	95.7	51.2	99.7	95.7
MSCM	✓	✓	80.6	79.8	88.4	90.7	80.5	98.9	98.8	71.5	99.8	98.9

4. Experimental setup and Result Analysis

A. Datasets and Training Labels

1)Datasetpreparation

This research relies on three publicly available datasets, i.e., TCGA GBM-LGG (Tomczak et al., 2015), CPTAC (Verdugo et al., 2022), and IvYGAP (Eberhart and Bar, 2020). In all datasets, entire slide images that have low image quality or no annotations are removed, and it is based on the preprocessing protocol in (Lu et al., 2021). There are a total of 3,578 WSIs in 1,054 cases.
In line with (Nguyen et al., 2023), both TCGA GBM-LGG and CPTAC data sets are combined to create a single meta-data set, which is utilized in the training and internal validation of the model. To be clear, this combined dataset is in this paper known as the TCGA GBM-LGG + CPTAC dataset. IvYGAP dataset is only used as an independent external validation dataset. Further information about preprocessing of datasets and inclusion criteria is also given in Section I of the supplementary material.

2) Training labels

Directly obtained ground-truth labels of the molecular markers and histological features are the three datasets. Depending on the new classification criteria proposed by WHO (2021) (Louis, 2021), the following labels of glioma subtypes are assigned to every case:
1.Grade 4 glioblastoma, which is IDH wild type;
2.Oligodendroglioma, which is defined as IDH mutant; 1p/19q co-deletion;
3.Grade 4 astrocytoma, which is considered IDH mutant, 1p/19q non-co- deleted with CDKN homozygous deletion or necrosis/microvascular proliferation (NMP);

4.All that remains are low-grade astrocytoma.

The diagnostic workflow of the used glioma classification according to WHO 2021 criteria is presented in Fig.1 of the supplementary material.

B. Implementation Details

The suggested MSCM is also trained on the TCGA GBM-LGG + CPTAC data with 50 epochs, a batch size of 6 and an initial learning rate of 0.003. Regularization is done with the Adam optimizer (Kingma and Ba, 2014). All hyper parameters are optimized depending on the performance of the internal validation set. Table I of the supplementary material contains a full list of settings of hyper parameters.

It is performed on a workstation based on AMD EPYC 7H12 with (1.70 GHz) CPU, 1 TB RAM, and eight Tesla V100 GPUs by NVIDIA. PyTorch is used to implement the model in Python.

C. Performance Evaluation

1) Glioma Classification

We benchmark the proposed MSCM framework on thirteen state-of-the-art (SOTA) models that comprise five popular image classification frameworks (AlexNet, ResNet-18, InceptionNet, MnasNet, and DenseNet-121), three state-of-the-art multiple instance learning (MIL) frameworks (ABMIL, TransMIL, and CLAM), two multi-modal learning frameworks (MCAT and CMTA), and three more recent frameworks that are specifically designed to classify glioma according to the WHO 2021 Importantly, the method of DeepMO-Glioma is the conference variant of our technique.

A fair and consistent evaluation of all comparison methods is achieved through the initiation of all comparison methods with unified glioma classification labels, which combine molecular markers with histological data. Performance is assessed using five standard metrics, accuracy, sensitivity, specificity, AUC, and F1-score, and micro-averaging is used on all metrics.

As it can be observed in Table 1, the proposed MSCM model records the highest overall results in the TCGA GBM-LGG + CPTAC dataset when compared with all the methods competing with it, by a decisive margin. Specifically, MSCM has an accuracy of 80.6, sensitivity of 79.8, AUC of 90.7 and F1-score of 80.5, which are at least 5-7 percentage points better than the best competing methods on these counts. These results support the fact that combining histology and molecular information to be jointly modeled in the same learning structure is effective.

This can be attributed to three important design decisions that have led to the observed improvements in performance:the inferred multi-scale disentangling module, a complementary tissue-and cellular-scale WSI characteristics captor;

1. the CPLC-Graph network, an explicit model of intrinsic correlations between molecular markers;
2. the proposed dynamic confidence constraint (DCC) loss and CMG-Modu training approach, which allows the cross-modal interaction between histology and molecular predictions to be effective. The role of each component is further evaluated in ablation study given in Section IV.D.

Moreover, the Fig.3 shows the ROC curves of glioma classification, according to which the proposed method prevails over the rival models throughout the operating point range, which proves its strong discriminative ability.

2) Molecular Marker and Histology Prediction Performance

In addition to the classification of glioma, we also test the performance of MSCM on auxiliary prediction tests such as IDH mutation, 1p/19q co-deletion, CDKN homozygous deletion, and necrosis/microvascular proliferation (NMP). The quantitative data are presented in Table II.

The suggested approach obtains 97.6, 94.3, 80.1, and 99.2 as the AUC values of IDH mutation, 1p/19q co-deletion, CDKN HOMDEL, and NMP prediction, respectively, which are significantly higher than all the comparison models. These findings prove the high ability of MSCM to deduce both molecular and histological features out of WSIs.

It is worth pointing out that the current multi-modal procedures like MCAT and CMTA uses the molecular markers as input and thus fails to carry out autonomous molecular marker prediction. Conversely, MSCM can be used to predict these markers with high accuracy based on only histology, which points to its usefulness in clinical practice, where molecular information is not always available.

Fig. 3 shows the corresponding ROC curves of all the auxiliary tasks that again confirm the strength and reliability of the proposed framework in various prediction objectives.

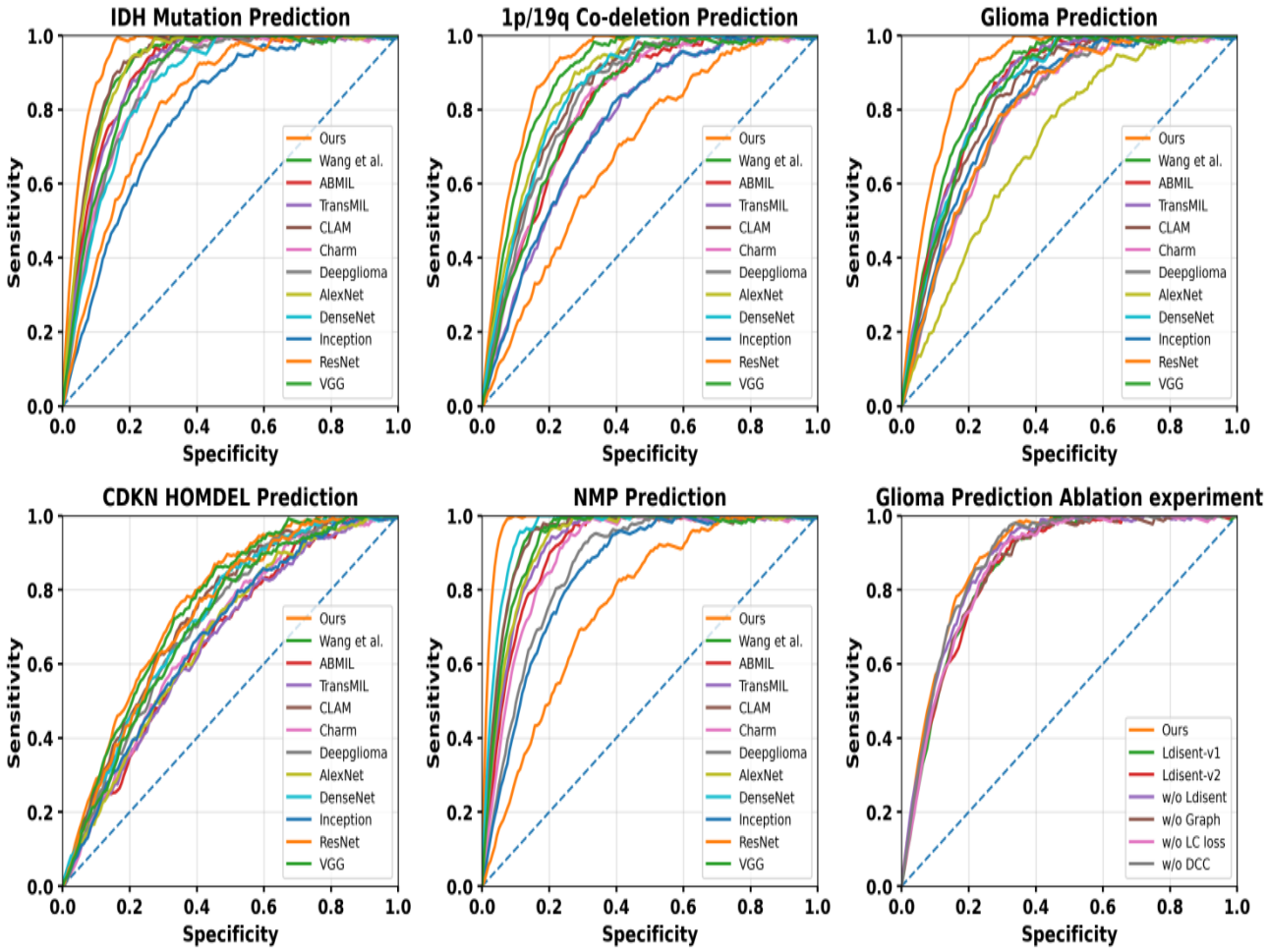

Fig 3. ROCs of our model, comparison and ablation models for predicting IDH, 1p/19q, CDKN, NMP and Glioma

3) Generalizability Evaluation

In order to evaluate model robustness and ability to generalize, we also do external validation tests on the IvYGAP data that do not perform fine-tuning. The findings are provided in the right hand side of Table I.

MSCM is shown to perform very well in generalization and has an accuracy of 98.9% and an AUC and F1-score of 99.8 and 98.9 respectively, which is significantly higher than the state-of-the-art approaches. These findings demonstrate high resistance to domain shift and heterogeneity of data sets, which highlights the trustworthiness of the presented framework in the implementation of the independent clinical setting.

Table II. Mean values (in %) for predicting molecular markers and histology on the TCGA GBM-LGG + CPTAC dataset

Task	Metrics	MSCM	Wang et al.	ABMIL	TransMIL	CLAM	Charm	Deepglioma	MCAT	CMTA	AlexNet	DenseNet	Inception	ResNet	VGG
IDH	Acc.	91.9	84.5	80.9	84.2	84.2	82.6	80.9	–	–	83.6	74.0	75.7	68.4	83.2
	Sen.	86.2	78.5	76.0	84.3	78.5	82.6	76.9	–	–	79.3	92.6	79.3	39.7	81.0
	Spec.	95.1	88.5	84.2	84.2	88.0	82.5	83.6	–	–	86.3	61.7	73.2	87.4	84.7
	AUC	97.6	92.1	90.5	89.7	92.7	87.8	87.3	–	–	91.9	86.5	78.9	82.0	88.9
	F1	89.3	80.2	76.0	81.0	79.8	79.1	76.2	–	–	79.3	73.9	72.2	50.0	79.4
1p/19q	Acc.	84.8	82.2	79.0	78.3	78.6	73.7	79.9	–	–	81.6	78.6	74.3	72.0	72.0
	Sen.	85.9	63.4	46.8	40.2	63.4	76.8	72.0	–	–	52.4	57.3	62.2	48.3	84.1

	Spec.	84.1	89.2	87.1	92.3	84.2	72.5	82.9	–	–	92.3	86.5	78.8	78.9	67.6
	AUC	94.3	89.1	81.6	76.4	85.1	82.0	83.7	–	–	86.8	85.4	76.6	69.0	81.5
	F1	74.6	65.8	47.3	50.0	61.5	61.2	65.9	–	–	60.6	59.1	56.7	43.9	61.9
CDKN	Acc.	76.9	71.1	63.5	63.8	65.1	64.1	62.2	–	–	62.8	61.5	60.5	71.6	64.8
	Sen.	81.4	79.4	59.0	68.8	81.9	58.1	46.3	–	–	75.6	50.6	41.3	48.3	58.8
	Spec.	72.8	61.8	69.7	58.3	46.5	70.8	79.9	–	–	48.6	73.6	81.9	78.4	71.5
	AUC	80.1	75.0	66.2	65.7	73.6	67.8	70.7	–	–	66.8	71.8	67.5	72.8	70.7
	F1	78.2	74.3	65.0	66.7	71.2	63.1	56.3	–	–	68.2	58.1	52.4	43.6	63.7
NMP	Acc.	96.8	90.1	82.4	87.5	86.2	84.5	80.9	86.0	86.6	85.2	89.8	78.3	69.1	81.6
	Sen.	96.1	88.0	93.6	87.4	82.0	82.6	82.6	75.9	77.8	80.2	90.4	73.7	68.9	71.9
	Spec.	97.4	92.7	69.4	87.6	91.2	86.9	78.8	90.7	90.7	91.2	89.1	83.9	69.3	93.4
	AUC	99.2	94.8	91.3	92.1	94.7	90.0	86.5	95.7	96.7	92.8	95.9	84.6	76.0	93.6
	F1	97.0	90.7	85.1	88.5	86.7	85.4	82.6	77.4	78.5	85.6	90.7	78.8	71.0	81.0

Table III. Subgroup analysis (in %) by WSI material on molecular markers, histology prediction, and glioma classification

Material	Task	Acc.	Sen.	Spec.	AUC	F1-score
FFPE	Diag	82.1	82.1	86.4	88.1	82.1
	IDH	90.1	89.2	90.9	96.1	89.8
	1p19q	80.1	96.6	69.6	92.2	79.2
	CDKN	80.1	76.5	83.1	82.8	77.6
	Grade	88.7	77.9	97.6	98.4	86.2
Frozen	Diag	75.2	75.2	78.7	89.2	75.2
	IDH	87.6	72.3	94.3	93.1	78.2
	1p19q	81.0	83.5	87.7	83.4	70.8
	CDKN	65.4	76.1	59.2	66.6	72.5
	Grade	78.4	66.7	100.0	98.1	80.0

5) Subgroup Analysis

In order to further assess robustness in various WSI materials, we subgroup-analyze frozen and FFPE sections in terms of glioma classification, molecular marker prediction and histology prediction. Table 3 provides quantitative results. Although some differences in performance are observed between frozen and FFPE sections, the overall disparities are relatively small. Specifically, glioma classification achieved values of AUC of 88.1% and 89.2% for FFPE and frozen sections respectively. These results indicate that the given method is applicable to a wide range of tissue preparation protocols and, therefore, should be used in a variety of clinical procedures.

D. Ablation Study

We also perform extensive ablation experiments to study the role of each of the key components in the proposed MSCM framework, such as the multi-scale disentanglement loss, the CPLC-Graph network, the label-correlation (LC) loss, the dynamic confidence constraint (DCC) loss, and the CMG-Modu training strategy. The training protocol of all the ablated versions is similar to the full model. Table IV shows quantitative findings on the TCGA GBM-LGG + CPTAC (internal) and IvYGAP (external) validation datasets, and Fig. 3 and Fig. 4 provide the corresponding ROC curves.

1) Multi-Scale Disentanglement Effect

In order to measure the effectiveness of the proposed multi-scale disentanglement module, we use three ablations on the loss L_{disent}:

1. Removing the constraint between modality-independent features of molecular markers and histology (L_{disent}-v1);

2. Removing the constraint between independent and shared features within each modality (L_{disent}-v2); and
3. Completely removing the disentanglement loss.
As in Table IV, each of the three variants exhibits a decrease in performance on the internal and external datasets relative to the full model. Specifically, the elimination of the entire disentanglement loss decreases the internal AUC from 92.6 to 90.3, and F1-score from 83.2 to 70.5, and proves that successful disentanglement is imperative in learning task-relevant and complementary representations. The same tendencies can be noticed in the external dataset, which proves the importance of the disentanglement module in enhancing generalization.

2) Effect of the CPLC-Graph Network
In order to gauge the contribution of the proposed CPLC-Graph network, graph modelling is turned-off by setting the graph balancing weight to zero. As Table IV shows, the removal of CPLC-Graph network results in a clear deterioration in performance across the evaluated metrics. In particularly, the internal accuracy and sensitivity decrease from 83.4% and 83.1% to 75.8% and 75.8%, respectively, while AUC decreases from 92.6% to 89.2%. It is also found that there is a drop in performance on the external dataset, which implies that modelling inherent associations between molecular indicators is crucial to solid glioma characterization. Fig.3 further supports these observations by the comparisons with ROCs.

Table IV. Ablation study on the multi-modal disentanglement loss, CPLC-Graph network, LC loss, and DCC loss on the TCGA GBM-LGG + CPTAC (internal) and IvYGAP (external) validation datasets

Ablation	**Internal Dataset**					**External Dataset**				
	Acc.	Sen.	Spec.	AUC	F1-score	Acc.	Sen.	Spec.	AUC	F1-score
Ldisent -v1	71.4	71.4	87.9	88.9	71.3	96.8	96.8	61.2	99.8	96.8
Ldisent -v2	73.2	73.2	85.6	89.1	73.1	94.9	94.9	72.8	99.4	94.9
w/o Ldisent	70.6	70.6	85.9	90.3	70.5	96.1	96.1	60.8	99.5	96.1
w/o CPLC-Graph	75.8	75.8	83.7	89.2	75.7	95.9	95.9	78.4	99.7	95.9
w/o LLC	68.9	68.9	89.2	89.4	68.8	94.3	94.3	95.1	99.1	94.3
w/o LDCC	73.9	73.9	92.1	91.2	73.8	95.4	95.4	83.2	98.6	95.4
MSCM	83.4	83.1	88.9	92.6	83.2	98.9	98.9	71.5	99.8	98.9

3) Effect of the Label-Correlation (LC) Loss
We also assess the effect of the LC loss by eliminating it through training. As Table IV shows, removing the LC loss leads to a significant deterioration in performance, where internal accuracy and internal sensitivity have gone down to 68.9% and 68.9%, and AUC has reduced to 89.4% as compared to 83.4% and 83.1% .These findings indicate that the LC loss improves consistency between learned feature similarities and biological co-occurrence statistics, thereby enhancing discriminative performance. Molecular marker and histology prediction tasks also exhibit similar degradation in performance (see Table 5), which shows the overall utility of the LC loss.

4) Effect of the Dynamic Confidence Constraint (DCC) Loss
The relevance of the proposed DCC loss is assessed by eliminating it out of the entire model. Table 4 indicates that the internal accuracy goes down to 73.9% and AUC goes down to 91.2%. The decrease in performance is also observed in the external dataset. Furthermore, elimination of the DCC loss results in significant decreases in the AUC of the auxiliary prediction tasks, as shown in Table V, and poor ROC performance in Fig. 3. These findings support the role of the DCC loss in strengthening biologically significant cross-modal histology-molecular marker interactions.

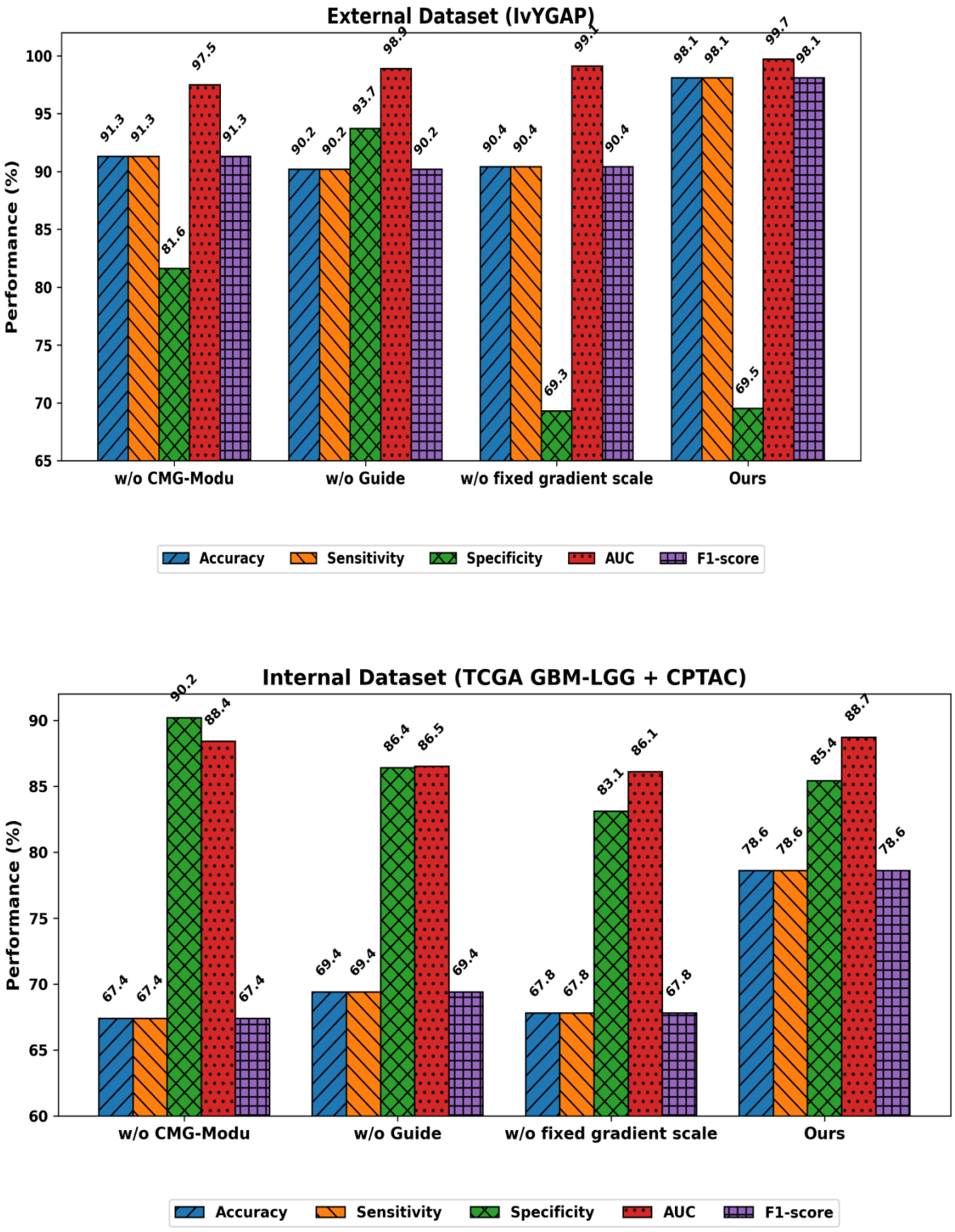

Fig. 4. Ablation study of CMG-Modu training strategy on the glioma classification task over internal and external validation datasets.

5) CMG-Modu Training Strategy Effect

Lastly, we determine the performance of the proposed CMG-Modu training strategy by comparing the complete model with various optimization strategies with modifications, such as eliminating gradient scaling, eliminating histology-guided modulation and eliminating CMG-Modu altogether. As demonstrated in Fig.4, all the altered strategies under discussion are worse than the entire MSCM framework both on internal and external data. This supports the use of dynamically regulated gradient flow guided by diagnostic cues as a promising strategy for reducing modality bias and improving optimization stability.

Table V. Ablation study on the multi-modal disentanglement loss, CPLC-Graph network, LC loss, DCC loss, and CMG-Modu training strategy on auxiliary tasks over the TCGA GBM-LGG + CPTAC dataset

Metrics	MSCM	Ldisent-v1	Ldisent-v2	w/o Ldisent	w/o Graph	w/o LC loss	w/o DCC	w/o CMG-Modu	w/o Guide	w/o fixed grad scale
Acc.	91.9	80.6	90.4	80.9	81.6	82.6	78.9	83.2	82.6	85.5
Sen.	86.2	93.4	68.8	81.0	81.0	82.6	71.1	80.2	87.6	89.3
Spec.	95.1	72.1	90.9	80.9	82.0	82.5	84.2	85.2	79.2	83.1
AUC	97.6	89.0	92.7	87.8	87.2	90.0	86.1	90.5	88.2	93.5
F1-score	89.3	79.3	76.8	77.2	77.8	79.1	72.9	79.2	80.0	83.1

1p/19q Co-deletion Prediction

Metrics	MSCM	Ldisent-v1	Ldisent-v2	w/o Ldisent	w/o Graph	w/o LC loss	w/o DCC	w/o CMG-Modu	w/o Guide	w/o fixed grad scale
Acc.	84.8	74.3	73.4	76.0	73.7	75.7	80.9	78.3	74.0	79.9
Sen.	85.9	82.9	84.1	51.2	72.0	85.4	84.1	57.3	74.4	59.8
Spec.	84.1	71.2	69.4	85.1	74.3	72.1	79.7	86.0	73.9	87.4
AUC	94.3	83.1	81.6	81.9	82.5	83.0	88.3	84.3	84.2	86.3
F1-score	74.6	63.6	63.0	53.5	59.6	65.4	70.4	58.7	60.7	61.6

CDKN HOMDEL Prediction

Metrics	MSCM	Ldisent-v1	Ldisent-v2	w/o Ldisent	w/o Graph	w/o LC loss	w/o DCC	w/o CMG-Modu	w/o Guide	w/o fixed grad scale
Acc.	76.9	68.8	61.2	66.8	68.4	68.1	67.4	70.4	69.4	72.0
Sen.	81.4	71.3	95.0	61.3	75.0	63.1	91.9	81.3	75.0	78.1
Spec.	72.8	66.0	23.6	72.9	61.1	73.6	40.3	58.3	63.2	65.3
AUC	80.1	72.8	72.7	75.5	71.8	75.2	75.1	73.8	74.6	73.8
F1-score	78.2	70.6	72.0	66.0	71.4	67.6	74.8	74.3	72.1	74.6

NMP Prediction

Metrics	MSCM	Ldisent-v1	Ldisent-v2	w/o Ldisent	w/o Graph	w/o LC loss	w/o DCC	w/o CMG-Modu	w/o Guide	w/o fixed grad scale
Acc.	96.8	84.2	87.8	88.2	87.2	92.1	84.9	90.1	87.5	91.1
Sen.	96.1	82.0	85.6	83.2	94.0	93.4	92.2	88.6	83.2	91.0
Spec.	97.4	86.9	90.5	94.2	78.8	90.5	75.9	92.0	92.7	91.2
AUC	99.2	93.1	93.3	93.8	93.1	96.3	93.0	94.7	93.5	96.2
F1-score	97.0	85.1	88.5	88.5	89.0	92.9	87.0	90.8	88.0	91.8

6) Overall Analysis

The findings in general in Table 4 show that all the elements of MSCM have a positive contribution to the performance, and the entire model performs optimally on both validation datasets. It is interesting to note that degradation of performance due to the removal of individual modules is moderate, which means that there is functional independence between the components to some extent. The same trends could be observed in the external dataset, which also confirms the success and external validity of the offered framework.

7. DISCUSSION

This paper introduces a multi-scale multi-task (MSCM) framework that is capable of classifying glioma and predicting the molecular markers and histologic features of the whole-slide images with one single framework. A substantial set of experimental findings on internal and external validation datasets prove that MSCM is consistently superior to state-of-the-art in all of the studied tasks, which proves the relevance of the combination of histology and molecular information as a single learning paradigm.

Multi-Scale and Multi-Task Learning Effectiveness

Among the main findings of the experimental results, there is that by jointly optimizing the performance on glioma patternization along with peripheral molecular markers and histology prediction tasks, considerable performance improvements are experienced. In comparison to models that are trained to classify glioma only, MSCM advantages the shared learning of features on related tasks, which promotes the network to learn the clinically significant patterns that are applicable across studies. The increase in performance with both internal and IvYGAP external datasets suggests that multi-task supervision is a good regularizer and it decreases the overfitting and enhances robustness.

Furthermore, the suggested multi-scale disentanglement module will provide MSCM to successfully seize complementary data of disparate magnifications of WSIs. Table 4 and Table 5, which are the results of the ablation, indicate that disentanglement constraints were removed or weakened, which caused some apparent degradation in the

glioma classification and auxiliary prediction tasks. The findings are that learning of discriminant and interpretable features of complex histopathological data would require explicit separation between modality-independent, modality-specific and shared representations.

Molecular Correlation Modeling Role

The suggested CPLC-Graph network is also another significant part of MSCM since it is used to model intrinsic correlations between molecular markers. In contrast to traditional methods in which molecular markers can be used as independent targets, MSCM directly models the biological connection of co-occurrence with each other in a graph-based message passing way. As it can be seen in the ablation experiment, impairment of the CPLC-Graph network leads to a stable decrease in performance in all the metrics, especially in glioma classification and prediction of molecular markers. This underlines the significance of putting into the design of models past biological knowledge to enhance predictive quality and integrity.

This conclusion is further proven by the effectiveness of the label-correlation (LC) loss. The LC loss makes biologically plausible predictions by matching learned similarities of features with co-occurrence statistics of molecular markers. The experimented decline in the performance by the removal of the LC loss ensures that modeling label dependencies explicitly makes a significant contribution to the enhanced classification and auxiliary task performances.

Significance of Cross-Modal Interaction

The suggested dynamic confidence constraint (DCC) loss and CMG-Modu training plan is very important in facilitating efficient cross-modal communication among histology and molecular predictions. In contrast to the case of static fusion strategies, DCC dynamically directs the model to concentrate on those spatial areas that are informative to both modalities. The obtained results of the ablation suggest that the deletion of the DCC loss results in stable decreases in the accuracy, AUC, and F1-score of the glioma classification and auxiliary tasks, which proves the importance of the DCC loss.

In the same manner, CMG-Modu training plan enhances the stability of optimisation by balancing the gradient flow using prevailing diagnostic signals. The problem of worse performance of modified training strategies attests to the fact that adaptive gradient modulation is essential to maintain balance between the efforts of the modalities and avoid any bias toward one of them. Combined, these elements allow MSCM to learn cross-modal representations which are coherent and clinically meaningful.

Generalizability and Clinical Relevance

It is also interesting to note that MSCM has a high generalization performance on the IvYGAP external data without fine-tuning. Although the data distribution differs and the composition of classes varies, MSCM has a high level of performance in comparison to the competing methods, which reflects the strength of its generalizability. The subgroup analysis also supports the fact that MSCM is very consistent between different WSI materials, such as FFPE and frozen sections, which indicates that MSCM can be used in a wide range of clinical contexts.

Along with this, the results of visualization prove that MSCM generates interpretable decision maps that inter-operate with the known diagnostic guidelines, including the robust correlation between NMP-positive histology with IDH wildtype status in glioblastoma. These results indicate that molecular marker prediction combined with histological analysis can not only be used to predict better but also to increase the model interpretability, which is a prerequisite of its use in clinical practice.

Future Research

Despite the good performance experienced by MSCM in various tasks, there are a number of limitations. First, the external validation dataset has less subtypes of glioma than the internal dataset and this could affect the absolute performance measures. Future directions will involve validation in multi-centers with different distributions of classes. Second, although the present framework is based on molecular markers related to glioma, it would be beneficial to apply the method to other types of tumors and molecular profiles, which would prove its broad applicability.

8. CONCLUSION

This paper offered a single multi-scale and multi-task model, MSCM, to classify glioma and assist in predicting molecular markers and histological signs directly on whole-slide images. MSCM has been shown to provide clinically meaningful representations and enhance the diagnostic performance because histology and molecular information are modeled jointly. Consecutive experiments on the TCGA GBM-LGG + CPTAC and IvYGAP datasets indicate that MSCM consistently accuracy, AUC, and F1-score than its state-of-the-art counterparts, and has high generalization capacity with respect to datasets and tissue preparation protocols. All of the proposed multi-scale disentanglement module, CPLC-Graph network, dynamic confidence constraint loss, and CMG-Modu training strategy yield performance improvements, which are confirmed by extensive ablation experiments. Also, the results of visualization indicate that MSCM yields interpretable predictions in accordance with the accepted diagnostic criteria. Altogether, MSCM provides a clinically applicable and powerful system of integrated glioma diagnosis based on histopathological images.

REFERENCES

1. D. Komura and S. Ishikawa, “Machine learning methods for histopathological image analysis,” Comput. Struct. Biotechnol. J., vol. 16, pp. 34–42, 2018.
2. A. Echle et al., “Deep learning in cancer pathology: A new generation of clinical biomarkers,” Br. J. Cancer, vol. 124, no. 4, pp. 686–696, 2021.
3. N. Coudray et al., “Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning,” Nat. Med., vol. 24, pp. 1559–1567, 2018.
4. T. J. Fuchs and J. M. Buhmann, “Computational pathology: Challenges and promises for tissue analysis,” Comput. Med. Imaging Graph., vol. 35, no. 7–8, pp. 515–530, 2011.
5. M. Ilse, J. M. Tomczak, and M. Welling, “Attention-based deep multiple instance learning,” in Proc. ICML, 2018, pp. 2127–2136.
6. J. Campanella et al., “Clinical-grade computational pathology using weakly supervised deep learning on whole slide images,” Nat. Med., vol. 25, pp. 1301–1309, 2019.
7. A. B. Tosun, C. Gunduz-Demir, “Graph-based representations for object detection and classification in histopathology,” IEEE Trans. Med. Imaging, vol. 30, no. 10, pp. 1853–1862, 2011.
8. L. Hou et al., “Patch-based convolutional neural network for whole slide tissue image classification,” in Proc. CVPR, 2016, pp. 2424–2433.
9. Y. Li et al., “Multi-scale attention networks for histopathology image classification,” Med. Image Anal., vol. 58, 2019.
10. G. Litjens et al., “A survey on deep learning in medical image analysis,” Med. Image Anal., vol. 42, pp. 60–88, 2017.
11. H. Kather et al., “Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer,” Nat. Med., vol. 25, pp. 1054–1056, 2019.
12. M. Lu et al., “Data-efficient and weakly supervised computational pathology on whole-slide images,” Nat. Biomed. Eng., vol. 5, pp. 555–570, 2021.
13. S. Graham et al., “Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images,” Med. Image Anal., vol. 58, 2019.
14. Y. Xu et al., “Large-scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features,” BMC Bioinformatics, vol. 18, no. 1, 2017.
15. J. P. Platt, “Probabilistic outputs for support vector machines,” in Adv. Large Margin Classifiers, MIT Press, 1999.
16. R. K. Srivastava, K. Greff, and J. Schmidhuber, “Highway networks,” in Proc. ICML, 2015.
17. A. Vaswani et al., “Attention is all you need,” in Proc. NeurIPS, 2017.
18. Z. Li et al., “Dual-stream multiple instance learning network for whole slide image classification,” IEEE Trans. Med. Imaging, vol. 39, no. 8, pp. 2643–2655, 2020.
19. S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, 2010.
20. X. Wang et al., “Multi-task learning for histopathological image analysis,” IEEE Trans. Med. Imaging, vol. 38, no. 2, pp. 470–480, 2019.
21. K. He et al., “Deep residual learning for image recognition,” in Proc. CVPR, 2016.
22. Y. Wang et al., “Weakly supervised deep learning for whole slide lung cancer image analysis,” IEEE Trans. Cybern., vol. 50, no. 9, pp. 3950–3962, 2020.
23. J. Lu et al., “Learning under label noise for histopathology image analysis,” Med. Image Anal., vol. 66, 2020.
24. A. Shaban et al., “Context-aware convolutional neural networks for cancer detection,” IEEE Trans. Med. Imaging, vol. 38, no. 10, pp. 2394–2404, 2019.
25. C. Chen et al., “Pathomic fusion: An integrated framework for fusing histopathology and genomic features,” IEEE Trans. Med. Imaging, vol. 41, no. 4, pp. 757–770, 2022.
26. Y. Fu et al., “Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis,” Nat. Cancer, vol. 1, pp. 800–810, 2020.
27. S. Mobadersany et al., “Predicting cancer outcomes from histology and genomics using deep learning,” Proc. Natl. Acad. Sci., vol. 115, no. 13, pp. E2970–E2979, 2018.
28. J. Xu et al., “Graph convolutional networks for cancer histology classification,” Med. Image Anal., vol. 70, 2021.
29. T. Zhou et al., “Learning deep features for molecular subtype prediction from histopathology,” Bioinformatics, vol. 36, no. 8, pp. 2421–2428, 2020.
30. L. Shen et al., “Deep learning to improve breast cancer detection on screening mammography,” Radiology, vol. 294, no. 2, pp. 293–302, 2020.
31. A. Raghu et al., “Transfusion: Understanding transfer learning for medical imaging,” NeurIPS, 2019.
32. J. Deng et al., “ImageNet: A large-scale hierarchical image database,” in Proc. CVPR, 2009.
33. H. K. Kim et al., “Molecular classification of gliomas,” Acta Neuropathol., vol. 131, pp. 803–820, 2016.
34. T. Ceccarelli et al., “Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma,” Cell, vol. 164, no. 3, pp. 550–563, 2016.
35. D. Louis et al., “The 2016 World Health Organization classification of tumors of the central nervous system,” Acta Neuropathol., vol. 131, pp. 803–820, 2016.
36. D. Louis et al., “WHO classification of tumours of the central nervous system,” 5th ed., IARC, 2021.
37. S. Bakas et al., “Advancing the cancer genome atlas glioma MRI collections,” Sci. Data, vol. 4, 2017.

38. F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in Proc. CVPR, 2017.
39. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, 2015.
40. M. Abadi et al., “TensorFlow: A system for large-scale machine learning,” in Proc. OSDI, 2016.
41. A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in Proc. NeurIPS, 2019.
42. J. Pearl, “Causality: Models, reasoning, and inference,” Cambridge Univ. Press, 2009.
43. J. Hu et al., “Squeeze-and-excitation networks,” in Proc. CVPR, 2018.
44. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. ICLR, 2015.
45. E. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, Basic Books, 2019.