



Data and text mining

Lung cancer subtype diagnosis using weakly-paired multi-omics data

Xingze Wang^{1,2}, Guoxian Yu ^{1,2,*}, Jun Wang ^{2,*}, Azlan Mohd Zain³ and Wei Guo^{1,2}

¹School of Software, Shandong University, Ji'nan 250100, China, ²SDU-NTU Joint Centre for AI Research, Shandong University, Ji'nan 250100, China and ³Big Data Centre, University Teknologi Malaysia, Skudai 81310, Malaysia

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on July 14, 2022; revised on August 30, 2022; editorial decision on September 18, 2022; accepted on September 19, 2022

Abstract

Motivation: Cancer subtype diagnosis is crucial for its precise treatment and different subtypes need different therapies. Although the diagnosis can be greatly improved by fusing multiomics data, most fusion solutions depend on paired omics data, which are actually *weakly paired*, with different omics views missing for different samples. Incomplete multiview learning-based solutions can alleviate this issue but are still far from satisfactory because they: (i) mainly focus on shared information while ignore the important individuality of multiomics data and (ii) cannot pick out interpretable features for precise diagnosis.

Results: We introduce an interpretable and flexible solution (LungDWM) for **Lung** cancer subtype **D**agnosis using **W**eakly paired **M**ultiomics data. LungDWM first builds an attention-based encoder for each omics to pick out important diagnostic features and extract shared and complementary information across omics. Next, it proposes an individual loss to jointly extract the specific information of each omics and performs generative adversarial learning to impute missing omics of samples using extracted features. After that, it fuses the extracted and imputed features to diagnose cancer subtypes. Experiments on benchmark datasets show that LungDWM achieves a better performance than recent competitive methods, and has a high authenticity and good interpretability.

Availability and implementation: The code is available at <http://www.sdu-idea.cn/codes.php?name=LungDWM>.

Contact: guoxian85@gmail.com or kingjun@sdu.edu.cn

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Lung cancer is the leading cause of worldwide cancer death and has the highest incidence rate among different cancer types (Howlader *et al.*, 2020). Non-small cell lung cancer (NSCLC) is the most typical lung cancer, it accounts for 85% of lung cancer patients, and small cell lung cancer takes the other 15%, the 5-year survival rate of NSCLC patients is only 13% (Zappa and Mousa, 2016). NSCLC can be further categorized into three major histopathological subtypes: 45–50% Adenocarcinoma (LUAD), 30–35% squamous cell carcinoma (LUSC) and 5–10% large cell (undifferentiated) carcinoma, which require different treatments. For example, small cell carcinoma often needs chemotherapy due to poor surgical treatment; LUAD needs surgical treatment or targeted therapy for effective intervention; while LUSC has bleeding and fewer mutations, so anti-angiogenesis drugs and targeted therapies are often ineffective, but immunotherapy can achieve good prognostic effects. Therefore, the accurate diagnosis of lung cancer subtypes is of paramount importance.

Many techniques have been proposed to diagnose lung cancer subtypes, such as computed tomography (CT), pathological examination and so on (Howlader *et al.*, 2020). Among them, the histology image in the pathological examination is the golden standard for cancer malignancy and subtypes diagnosis. With the advance of sequencing technologies, liquid biopsy has become a non-invasive and effective way for early cancer diagnosis and targeted therapy (Crowley *et al.*, 2013). Besides, detection techniques at different biological levels have also been used, for example assessing single nucleotide variation, DNA methylation (Hao *et al.*, 2017) and miRNA expression quantification (Ahmed *et al.*, 2021).

With the surge of artificial intelligence (AI) techniques and multiomics data, *AI+omics*-based techniques have been explored for subtype diagnosis (Lehman and Wu, 2021; Menyhárt and Györfy, 2021). To name a few, Coudray *et al.* (2018) trained a deep CNN on tiled patches of whole-slide images collected from TCGA (Weinstein *et al.*, 2013) for lung cancer subtype diagnosis. Hao *et al.* (2017) applied the LASSO classifier to distinguish the tumor and normal tissues of four common cancers using genome-wide DNA

methylation data. These single-omics-based methods can only partially uncover the pathology, while complex cancers are jointly caused by multiple-level molecules. In addition, single-omics data are often noisy, incomplete and with low coverage, multiomics data can not only overcome these negative impacts but also provide a more holistic biological atlas.

More recent efforts fuse multiomics data by multiview learning (MVL) to improve the diagnosis performance. Existing MVL-based solutions mainly build on canonical correlation analysis, cotraining, matrix factorization and multiple kernel learning to integrate heterogeneous omics data (Gligorijević and Pržulj, 2015; Li *et al.*, 2018). However, they typically can only capture the shallow correlations among omics. Deep MVL-based methods can mine more complex correlations and thus have also been applied for cancer diagnosis and prognosis prediction (Yang *et al.*, 2021). For example, MDNNMD (Sun *et al.*, 2019) integrates multiomics data by a score-level fusion of the prediction results for breast cancer prognosis prediction. Mobadersany *et al.* (2018) merged the image and genomics data into a Cox proportional hazards model to predict patient outcomes. LungDIG (Wang *et al.*, 2021b) combines image and genomics data for interpretable lung cancer subtype diagnosis. However, due to the high cost of monitoring facilities (i.e. CT and gene test), invasive examinations (i.e. pathological biopsy and thoracentesis), legal and ethical constraints, multiomics data are mostly *weakly paired* (also termed as modal missing or incomplete multiomics data), with several omics of the same samples missing, which are ubiquitous in reality and break the prerequisite of completely paired omics data for canonical MVL solutions.

Incomplete MVL (iMVL)-based techniques have been studied to fuse weakly paired omics data (Li *et al.*, 2018). Yuan *et al.* (2012) proposed two iMVL methods, namely incomplete multisource feature learning (iMSF) and score completion (ScoreComp) to diagnose Alzheimer's disease. iMSF divides multiomics data into several learning tasks according to the omics availability and trains a unique classifier for each task, then learns the shared features across omics. ScoreComp separately trains a base classifier for each omics, then estimates missing prediction scores using other scores predicted by base classifiers. CG19 (Cheerla and Gevaert, 2019) first encodes the patient's multiomics data into vectors of the same dimension, maximizes the similarity of feature vectors from the same patient and minimizes those from different patients, then uses the Cox network

to predict the prognosis of cancer. CPM-GAN (Zhang *et al.*, 2022) imputes missing data through generative adversarial network (GAN; Goodfellow *et al.*, 2014), then maps different views into a shared representation for disease diagnosis. GIMPP (Arya and Saha, 2021) incorporates multiomics encoder networks and a bimodal attention mechanism to learn shared latent representations, and further uses GAN to generate missing data based on shared representations to diagnose cancer.

Those iMVL-based methods have confirmed the benefit of fusing incomplete multiomics data, but still face *three challenges*. First, most methods mainly pursue the shared/complementary features across omics but neglect the individuality of each omics (Arya and Saha, 2021; Cheerla and Gevaert, 2019; Zhang *et al.*, 2022). Given the variety of subtypes, the balance of individual and shared features of multiomics data enables the model to have a better diagnostic performance. Second, existing solutions rely on too stringent prerequisites to be applied in practice, such as excluding samples with missing omics (Arya and Saha, 2021; Wang *et al.*, 2021a), completing samples for at least one omics (Yuan *et al.*, 2012) and building models based on data availability (Wang *et al.*, 2020; Yuan *et al.*, 2012). Third, although deep iMVL methods often perform better than shallow ones (Cheerla and Gevaert, 2019; Rappoport and Shamir, 2019; Zhang *et al.*, 2022), their interpretability and authenticity remains to be improved, which prohibits their applications in evidence-based diagnosis.

To address these challenges, we propose an approach called LungDWM (**L**ung cancer subtype **D**iagnosis using **W**eakly paired **M**ultiomics data) and present the conceptual framework in Figure 1. LungDWM firstly trains an attention-based encoding network for each omics to extract the shared/complementary features across omics and to pick out key features for subtype diagnosis. To account for the variety of subtypes, it introduces an individual loss to extract the specific features of each omics. In addition, it designs a generative adversarial strategy to impute the missing omics using extracted features and thus enables flexible diagnosis. LungDWM finally fuses the extracted and imputed features to diagnose subtype. Experimental results on TCGA data show that LungDWM achieves a better diagnosis with good interpretability [Accuracy of 0.942, area under the receiver operating characteristics curve (AUROC) of 0.961, F1-Score of 0.937 and area under precision-recall curve (AUPRC) of 0.958] than competitive approaches (Arya and Saha,

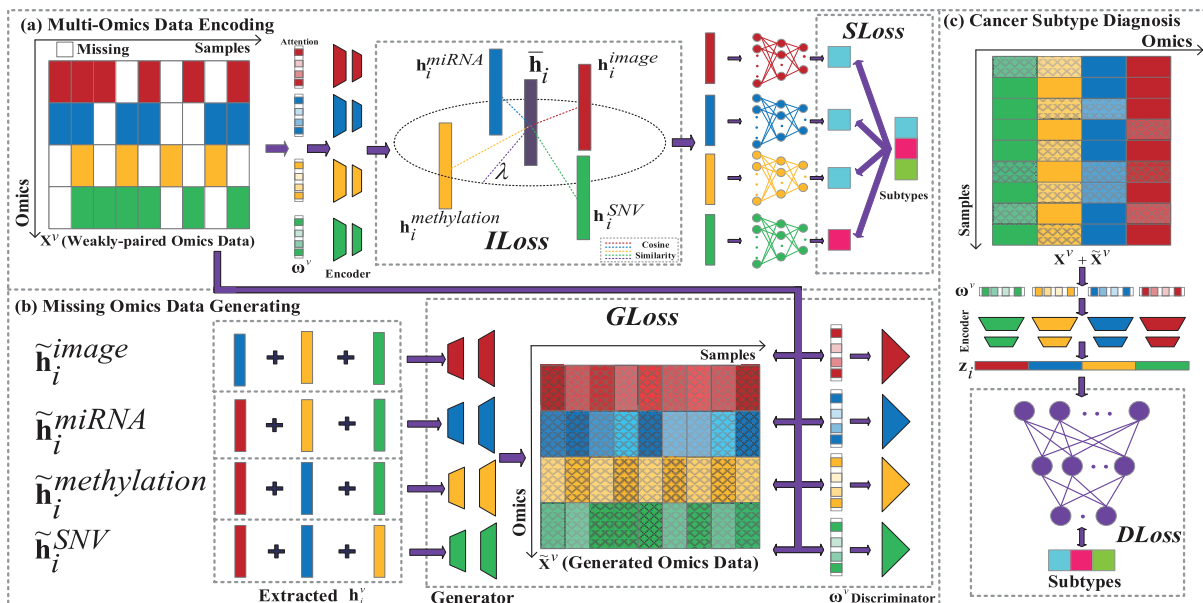


Fig. 1. Schema framework of LungDWM: (a) multiomics data encoding module uses an attention-based encoder to extract omics features h_i^v from weakly paired multiomics data X^V , and balances the shared and specific features in h_i^v by jointly optimizing the shared loss (SLoss) and individual loss (ILoss); (b) missing omics data generating module leverages the attention weights and available omics data to impute the missing omics data, and enhances the data integrity; (c) cancer subtype diagnosis module fuses extracted multiomics data to diagnose subtype by multilayer perceptron (MLP)

2021; Cheerla and Gevaert, 2019; Sun et al., 2019; Wang et al., 2021b; Yuan et al., 2012; Zhang et al., 2022), it also has a better performance on the Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) data (Su et al., 2020). LungDWM can aid pathologists to identify potential causal and therapeutic sites via attention weights learned from multiomics data.

2 Materials and methods

2.1 Overview and formulation

LungDWM diagnoses cancer subtypes by attention-based deep iMVL and GAN to integrate the weakly paired multiomics data. Figure 1 shows the basic workflow of our model. First, LungDWM uses an attention-based encoder to extract the omics features \mathbf{h}_i^v from weakly paired multiomics data \mathcal{X} , and leverages the joint optimization of shared loss (SLoss) and individual loss (ILoss) to balance the shared and specific features in \mathbf{h}_i^v . Second, it leverages the attention weights and available omics data to impute the missing omics data and enhances the data integrity. Finally, it fuses extracted multiomics data to diagnose subtype by multilayer perceptron. The mainly used symbols are listed in Table 1.

2.2 Multiomics data encoding module

To effectively integrate multiomics data, we build an attention-based encoder for each omics, which can not only maps heterogeneous features of different omics into a shared feature space but also helps these encoders not over-/under-fit to individual omics data. In addition, the encoder can alleviate the impact of missing and noisy features (Vincent et al., 2010). The encoding process is formulated as follows:

$$\mathbf{h}_i^v = f_{\Phi}^v(\omega^v \odot \mathbf{x}_i^v), \quad (1)$$

where $f_{\Phi}^v(\cdot)$ is the feature encoding network of the v -th omics parameterized with f_{Φ}^v , $\mathbf{h}_i^v \in \mathbb{R}^q$ is the encoded representation of \mathbf{x}_i^v , \odot is the element-wise multiplication operator. $\omega^v \in \mathbb{R}^{d_v}$ are the attention weights to pick out informative features, they are processed by *Softmax* to prevent the local optima problem caused by a too large weight of the significant position.

To enable the attention weights and encoded feature vectors with good authenticity and diversity, we optimize the feature encoding network by introducing a shared loss (SLoss) and an individual loss (ILoss). For the shared loss, we construct a feature evaluation network f_{Ω}^v on \mathbf{h}_i^v and quantify the loss based on the evaluated subtype given by f_{Ω}^v and the ground truth as:

$$S_{Loss} = \sum_{v=1}^V \frac{\sum_{i=1}^N \Lambda_i^v L_s(f_{\Omega}^v(\mathbf{h}_i^v), y_i)}{\sum_{i=1}^N \Lambda_i^v}, \quad (2)$$

where $L_s(f_{\Omega}^v(\mathbf{h}_i^v), y_i)$ is a loss function (cross entropy loss is used here) that measures the representation ability of \mathbf{h}_i^v , and $\Lambda_i^v \in \{0, 1\}$ indicates whether \mathbf{x}_i^v of the i -th patient is missing or not. SLoss aims to induce consistent predictions close to the ground truths, no matter the particular omics data are available or not. We use SLoss to jointly optimize the feature evaluation network and the attention-based encoder networks. By doing so, we can not only optimize the attention layers and encoder networks toward extracting

Table 1. Mainly used symbols

Notation	Description
$\mathcal{X} = \{\mathbf{X}^v\}_{v=1}^V$	A dataset with N samples and V types of omics data
$\mathbf{X}^v \in \mathbb{R}^{N \times d_v}$	The v -th omics data with d_v -dimensional features
$\mathbf{Y} \in \{1, \dots, s\}^N$	Subtypes of N patients
$\Lambda \in \mathbb{R}^{N \times V}$	Indicator matrix for weakly paired multi-omics data
$\mathbf{x}_i^v \in \mathbb{R}^{d_v}$	Feature vector of the i -th sample of \mathbf{X}^v
$\tilde{\mathbf{x}}_i^v \in \mathbb{R}^{d_v}$	Generated feature vector of \mathbf{x}_i^v
$\mathbf{h}_i^v \in \mathbb{R}^q$	q -Dimensional representation vector of \mathbf{x}_i^v

informative features from multi-omics data that contribute to subtype diagnosis but also improve the interpretation and authenticity.

SLoss mainly focuses on the shared/complementary features across omics and down-weights specific features of each omics. Given the variety among subtypes of the same cancer, it may override the diversity of cancer subtypes, which is crucial for precise cancer subtype diagnosis. The individuality and commonality of multiomics data should be jointly used for accurate diagnosis (Tan et al., 2021). Given that, we further propose the individual loss (ILoss) to rectify \mathbf{h}_i^v to maintain the diversity of subtypes by balancing individual and shared features as follows:

$$I_{Loss} = \sum_{v=1}^V \frac{\sum_{i=1}^N \Lambda_i^v Relu(\lambda - \cos(\mathbf{h}_i^v, \bar{\mathbf{h}}_i))}{\sum_{i=1}^N \Lambda_i^v}, \quad \bar{\mathbf{h}}_i = \frac{\sum_{v=1}^V \Lambda_i^v \mathbf{h}_i^v}{\sum_{v=1}^V \Lambda_i^v}, \quad (3)$$

where $\cos(\mathbf{h}_i^v, \bar{\mathbf{h}}_i)$ is the cosine similarity between \mathbf{h}_i^v and shared feature vector $\bar{\mathbf{h}}_i$. When the similarity is larger than the individual factor λ ($-1 \leq \lambda \leq 1$), \mathbf{h}_i^v are more similar to the shared ones, and thus avoid over-individualization; otherwise, \mathbf{h}_i^v under-weights specific features. Therefore, we use *Relu*(\cdot) activation function to transform this loss to 0. SLoss aims to extract shared and complementary information from multiomics data. Using only ILoss to learn specific features may make the feature encoder module paying more attention to noisy/specific features in each omics data, which are not conducive to the disease diagnosis. Through the joint optimization of SLoss and ILoss, LungDWM can extract the shared and specific features from multiomics data. In return, $\{\mathbf{h}_i^v\}_{v=1}^V$ maintains the diversity of cancer subtypes and enables an accurate diagnosis.

2.3 Missing omics data generating module

It is impractical and even infeasible to collect all omics data of the same patients. In practice, only one or two omics data of the same patient are available for the diagnosis. Therefore, multiomics data of cancer samples are *weakly paired*. While recent iMVL-based methods (Arya and Saha, 2021; Zhang et al., 2022) can impute missing omics data from other omics by GAN, which is more adaptive to diverse input distributions. But they focus on generating all features in the missing omics and thus have unnecessary losses. Here, we leverage GAN with attention weights obtained in the encoding module to only impute important omics features. In this way, LungDWM can leverage available omics data to make a flexible diagnosis and avoid the risk of using single-omics data alone.

A typical GAN consists of two subnets: a generative subnet $G(\cdot)$ that learns to generate missing omics features based on $\{\mathbf{h}_i^v\}_{v=1}^V$ of available omics, and a discriminative subnet $D(\cdot)$ that recognizes whether the features are from available omics or from $G(\cdot)$. We first induce the potential features of a patient based on \mathbf{h}_i^v of all other available omics as:

$$\tilde{\mathbf{h}}_i^v = \frac{\sum_{j \in \{1, \dots, V\}, j \neq v} \Lambda_j^v \mathbf{h}_i^j}{\sum_{j \in \{1, \dots, V\}, j \neq v} \Lambda_j^v}. \quad (4)$$

Next, we input $\tilde{\mathbf{h}}_i^v$ into G^v to impute the missing data as:

$$\tilde{\mathbf{x}}_i^v = G^v(\tilde{\mathbf{h}}_i^v). \quad (5)$$

To maintain the generation ability, the v -th omics data are excluded for inducing $\tilde{\mathbf{h}}_i^v$. $G^v(\cdot)$ is the generator to generate the v -th omics data, and $\tilde{\mathbf{x}}_i^v \in \mathbb{R}^{d_v}$ is the imputed feature vector.

Compared with traditional GAN, Wasserstein GAN-Gradient Penalty (WGAN-GP; Gulrajani et al., 2017) not only solves the training instability of GAN caused by the imbalance of training levels of generator and of discriminator but also ensures the diversity of generated data. So we adopt the Wasserstein distance to improve the generative ability of G^v . In addition, to focus on important features of generated omics data, we input the real and generated omics data weighted by attention parameters into the discriminator subnet D^v , and compute the distribution values of samples as:

$$P_r^v = \frac{\sum_{i=1}^N \Lambda_i^v D^v(\omega^v \odot \mathbf{x}_i^v)}{\sum_{i=1}^N \Lambda_i^v}, P_g^v = \frac{\sum_{i=1}^N \Lambda_i^v D^v(\omega^v \odot \tilde{\mathbf{x}}_i^v)}{\sum_{i=1}^N \Lambda_i^v}, \quad (6)$$

where $D^v(\cdot)$ is a discriminator that takes the generated sample $\tilde{\mathbf{x}}_i^v$ or the real sample \mathbf{x}_i^v of v -th omics as input, and its output is the distribution value of the input sample in the 1D feature space. P_r^v and P_g^v are the average distribution values of real and generated samples, respectively. In addition, to meet the Lipschitz condition in WGAN (Arjovsky *et al.*, 2017), we apply gradient penalty to the discriminator by randomly sampling as:

$$\hat{\mathbf{x}}_i^v = \epsilon \mathbf{x}_i^v + (1 - \epsilon) \tilde{\mathbf{x}}_i^v \quad (7)$$

$$L_{gp}^v = \frac{\sum_{i=1}^N \Lambda_i^v (\|\nabla_{\hat{\mathbf{x}}_i^v} D^v(\omega^v \odot \hat{\mathbf{x}}_i^v)\|_2 - 1)^2}{\sum_{i=1}^N \Lambda_i^v}, \quad (8)$$

where $\hat{\mathbf{x}}_i^v$ is uniformly sampled through ϵ ($0 \leq \epsilon \leq 1$) along the line between a pair of points from the real distribution \mathbf{x}_i^v and generative one $\tilde{\mathbf{x}}_i^v$. $\nabla_{\hat{\mathbf{x}}_i^v} D^v(\omega^v \odot \hat{\mathbf{x}}_i^v)$ is the gradient and L_{gp}^v is the average of gradient penalty loss. P_r^v and P_g^v may be negative, we use a piecewise function under the joint gradient penalty term to compute the discrimination loss and generation loss as follows:

$$L_G^v = \begin{cases} a^{-P_g^v}, & P_g^v > 0 \\ -P_g^v, & P_g^v \leq 0 \end{cases} \quad (9)$$

$$L_D^v = \begin{cases} a^{P_g^v - P_r^v} + L_{gp}^v, & P_g^v - P_r^v < 0 \\ P_g^v - P_r^v + L_{gp}^v, & P_g^v - P_r^v \geq 0 \end{cases}, \quad (10)$$

$a > 1$ is a scalar parameter to convert the negative loss into a positive one, and the constraints on L_{gp}^v guide discriminator to make the distribution value of generated data close to but not exceed that of real data during the optimization process. Finally, we use L_D^v and L_G^v of V omics data to obtain the adversarial loss of LungDWM in the generating module as:

$$GLoss = \sum_{v=1}^V (L_G^v + L_D^v). \quad (11)$$

Through the optimization of $GLoss$, we can gradually improve the generator subnet in the process of generation and adversarial. In addition, since the data used by discriminator are weighted by attention parameters, which can down-weight the less important features, our generator can more focus on important features and thus improve the diagnosis authenticity.

2.4 Cancer subtype diagnosis module

Based on extracted features \mathbf{h}_i^v and the generated ones $\tilde{\mathbf{x}}_i^v$ for a missing omics, we train the diagnosis network by fusing them as:

$$\mathbf{z}_i^v = \Lambda_i^v \mathbf{h}_i^v + (1 - \Lambda_i^v) f_{\Phi}^v(\omega^v \odot \tilde{\mathbf{x}}_i^v) \quad (12)$$

$$\mathbf{z}_i = [\mathbf{z}_i^1; \mathbf{z}_i^2; \dots; \mathbf{z}_i^V], \quad (13)$$

where $\mathbf{z}_i^v \in \mathbb{R}^q$ is the representational feature vector of the v -th omics that will be used for cancer subtype diagnosis for patient i . When \mathbf{x}_i^v is available, we directly use its representation \mathbf{h}_i^v for subsequent feature fusion; otherwise, we apply f_{Φ}^v on generated $\tilde{\mathbf{x}}_i^v$ to obtain its representation for follow-up fusion. $\mathbf{z}_i \in \mathbb{R}^{q \cdot V}$ is the concatenated features for cancer subtype diagnosis. Considering that \mathbf{h}_i^v are salient features weighted by the attention mechanism and $\tilde{\mathbf{x}}_i^v$ is generated by \mathbf{h}_i^j ($j \neq v$), we do not weight them here.

We then input \mathbf{z}_i into the cancer subtype diagnosis network to predict the cancer subtype and compute the diagnosis loss (DLoss) as:

$$DLoss = \frac{1}{N} \sum_{i=1}^N CE(f_{\Psi}(\mathbf{z}_i), y_i), \quad (14)$$

where $f_{\Psi}(\cdot)$ is the diagnose network parameterized by Ψ , $CE(f_{\Psi}(\mathbf{z}_i), y_i)$ is the cross entropy loss function.

To this end, we formulate the objective function of LungDWM as:

$$L = \min_{\Phi, \Omega, G, D, \Psi} SLoss + ILoss + GLoss + DLoss. \quad (15)$$

By optimizing these modules, the missing omics data can be more reliably imputed from the available ones, both the shared and specific features are extracted from multiomics. As such, LungDWM can capture and preserve the variety of cancer subtypes, and give a more accurate and authentic diagnosis of cancer subtypes.

3 Results and validation

We testify LungDWM on TCGA Lung cancer data and SARS-CoV-2 data, and then perform ablation experiments to study the key modules of LungDWM. We further evaluate the robustness of LungDWM and other compared methods under different settings of missing data, and investigate the authenticity of LungDWM for clinical diagnostic.

3.1 Results on lung cancer subtype diagnosis

We downloaded the TCGA Lung Cancer data, including single nucleotide variation, DNA methylation, miRNA and tissue whole slide image, to quantitatively evaluate the performance of LungDWM. We want to remark that our model can also fuse other types of omics data (i.e. mRNA expression and copy number variation). The cancer samples of small cell carcinoma and large cell carcinoma are relatively scarce, and we did not collect samples of these two subtypes from TCGA, so we only considered LUAD and LUSC subtype data of NSCLC for experiments. For a comprehensive and comparative evaluation, we take seven methods for comparison, including *vanilla MVL-based methods* [MDNNMD (Sun *et al.*, 2019) and LungDIG (Wang *et al.*, 2021b)] that only use well-paired omics data to diagnose cancer subtypes; *iMVL-based methods* [iMSF (Yuan *et al.*, 2012), ScoreComp (Yuan *et al.*, 2012), CG19 (Cheerla and Gevaert, 2019), CPM-GAN (Zhang *et al.*, 2022) and GIMPP (Arya and Saha, 2021)] that diagnose subtypes using weakly paired omics data. All compared methods were discussed in Section 1. For MDNNMD and LungDIG, we separately used k -Nearest Neighbor (k NN) and zero values to impute the missing omics data of individual patients, and termed the corresponding methods as MDNNMD- k NN and MDNNMD-zero, LungDIG- k NN and LungDIG-zero. The configurations of compared methods and details of used datasets are given in [Supplementary Section S1 of Supplementary file](#).

We utilize four canonical evaluation metrics: Accuracy, AUROC, F1-Score and AUPRC to evaluate the diagnosis results of compared methods. A larger value of these metrics indicates a better performance. We assess the statistical significance at the 95% level by paired t -test, use \bullet/\circ to indicate that LungDWM performs better/worse than the other method. [Table 2](#) reports the average and standard deviation of Accuracy, AUROC, F1-Score and AUPRC of each method on the Lung cancer dataset, where the best results are shown in bold face. From these results, we can observe the followings:

- i. iMVL-based methods are more effective for cancer subtype diagnosis than MVL-based ones on weakly paired omics data. LungDWM improves the Accuracy by 3.6%, AUROC by 3.6%, F1-Score by 3.2% and AUPRC by 3.9% to the best MVL-based method LungDIG-zero. Other deep iMVL methods also often have better results than MVL-based ones. That is because iMVL-based methods account for intrinsic incompleteness of

Table 2. Results of compared methods on diagnosing TCGA Lung cancer data

	Accuracy	AUROC	F1-Score	AUPRC
MDNNMD- <i>k</i> NN	0.862±0.024•	0.913±0.015•	0.861±0.030•	0.906±0.017•
MDNNMD-zero	0.881±0.022•	0.920±0.012•	0.878±0.030•	0.909±0.016•
LungDIG- <i>k</i> NN	0.899±0.013•	0.917±0.013•	0.892±0.017•	0.913±0.019•
LungDIG-zero	0.909±0.014•	0.927±0.012•	0.907±0.015•	0.922±0.016•
iMSF	0.896±0.011•	0.918±0.011•	0.889±0.015•	0.912±0.010•
ScoreComp	0.901±0.013•	0.913±0.015•	0.897±0.016•	0.916±0.016•
CG19	0.914±0.007•	0.930±0.009•	0.913±0.007•	0.930±0.010•
CPM-GAN	0.923±0.017•	0.944±0.015•	0.922±0.016•	0.939±0.016•
GIMPP	0.930±0.013•	0.952±0.014	0.924±0.013•	0.945±0.012•
LungDWM	0.942±0.011	0.961±0.011	0.937±0.012	0.958±0.014

Best results are shown in bold face.

- multiomics data, and simply imputing the missing data with zero or by *k*NN is not so effective.
- ii. Deep learning-based methods have a good potential in diagnosis than shallow ones. Compared with shallow iMSF and ScoreComp, deep iMVL-based methods manifest a better diagnostic performance. Compared with ScoreComp, LungDWM achieves the improvement of 3.0% in Accuracy, 4.6% in AUROC, 4.4% in F1-Score and 4.5% in AUPRC. This is because deep methods can better capture complex correlations of multiomics data.
 - iii. The attention-based deep models (i.e. LungDIG-*k*NN, LungDIG-Zero, GIMPP and LungDWM) can better handle the overfitting problem caused by the high-dimensional omics data than non-attention-based ones. LungDWM improves the non-attention-based method CPM-GAN in Accuracy by 2.1%, AUROC by 1.8%, F1-Score by 1.6% and AUPRC by 2.0%; it also has better results than other attention-based solutions. This supports that attention weights in LungDWM can better capture the significant features of multiomics data for authentic diagnosis.
 - iv. Introducing the attention weights into GAN makes the generated data more helpful for subtype diagnosis. Compared with CG19 without GAN, GAN-based iMVL methods (CPM-GAN, GIMPP and LungDWM) have a clear better performance. This is because GAN can generate missing omics using available ones, and is adaptive to different distributions. In addition, compared with other GAN-based methods, LungDWM improves the Accuracy by 1.2%, AUROC by 0.9%, F1-Score by 1.4% and AUPRC by 1.3% to the second best performer. This fact proves that the attention weights help LungDWM to more focus on important features, which are beneficial for cancer diagnosis using the generated omics data, rather than all features. As a result, the extracted features from generated omics are more helpful than other GAN-based solutions for cancer subtype diagnosis.
 - v. The balance of specific and shared features is essential for the precise diagnosis of cancer subtypes. Compared with the best results among deep iMVL methods (CG19, CPM-GAN and GIMPP), our LungDWM improves the Accuracy by 1.2%, AUROC by 0.9%, F1-Score by 1.4% and AUPRC by 1.3%. That is because other methods mainly focus on the shared/complementary features across omics, alike our SLoss in Equation (2) does. The ILoss can rectify the encoding network to explore and preserve the shared and specific features of multiomics data. For this advantage, the extracted features preserve the variety of subtypes and enable more accurate diagnosis. Our ablation study will further confirm this advantage.

Besides, we conducted experiments on SARS-CoV-2 dataset (Su et al., 2020) to study the generalization of our LungDWM. The results and analysis are provided in Supplementary Table S5 and Section S2 of Supplementary file. In summary, these results prove the effectiveness of LungDWM for precise cancer subtype diagnosis using weakly paired omics data.

3.2 Ablation study

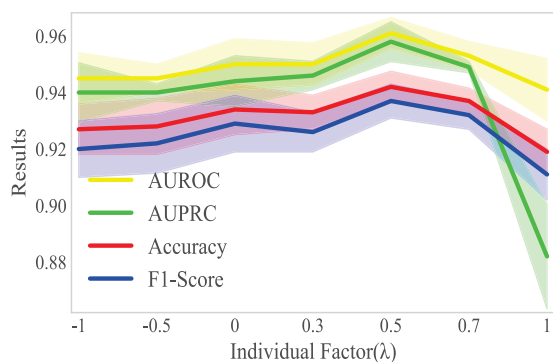
To study the contribution factors of LungDWM, we introduce five variants: LungDWM-w/oAtns, LungDWM-w/oGANs, LungDWM-GANs w/oAtns, LungDWM-w/oSLoss and LungDWM-w/oILoss, which separately disregard the attentions in multiomics encoders, GANs, attentions in GANs, SLoss and ILoss. Table 3 records the average results of LungDWM and its variants.

We see that LungDWM outperforms its variants by a distinct margin, which suggests that: (i) The attention mechanism helps the encoder module to better mine the significant features from multiomics data for cancer subtypes diagnosis, and LungDWM-w/oAtns has the largest performance drop. This proves the effectiveness of the attention mechanism in LungDWM on learning high-quality representations from omics data with a large number of missing and noisy sites. (ii) The GAN module can impute the missing omics data during the adversarial process and boost the fusion of weakly paired multiomics data. (iii) Adding the attention weights into GAN helps to focus on important features in the generated omics data that are helpful for diagnosis, and making the generated data are more similar with the real one, and thus enables a high-quality imputation of missing omics data. (iv) The leverage of SLoss and ILoss enables the encoder module to extract and preserve both shared and complementary features, and specific features of respective omics, which maintain the variety of cancer subtypes and improve the diagnosis authenticity and precision.

To further study the benefits of ILoss and its relation with SLoss, we vary the hyperparameter λ ($-1 \leq \lambda \leq 1$) in Equation (4), and reveal the results in Figure 2. We observe that the balance of shared/complementary and specific features of multiomics data ($\lambda = 0.5$) can make the fused features more diverse, which improves the performance by 2.5% in Accuracy, 2.1% in AUROC, 2.8% in F1-Score and 8.6% in AUPRC than LungDWM with a large $\lambda = 1$. LungDWM also has an improvement of 1.6% on Accuracy, 1.6% on AUROC, 1.8% on F1-Score and 1.9% on AUPRC with a small $\lambda = -1$. This is because a too larger or smaller λ is not conducive to a more differential fusion of multiomics data. An extreme large $\lambda \approx 1$ forces these encoders to more focus on the shared/complementary features across omics, and under-weight the specific ones; while $\lambda \approx -1$ gives the opposite. Both extreme cases have a compromised performance. Compared with the over-individualization of multiomics data, over-emphasis on shared/complementary features across omics results in a more severe performance drop (especially on AUPRC), the latter loses a lot of specific information and leads to a severe overfitting problem. This pattern confirms the necessity and effectiveness of extracting shared and specific features across omics for preserving the diversity of cancer subtypes.

Table 3. Results of LungDWM and its variants on TCGA Lung cancer data

Variant	Accuracy	AUROC	F1-Score	AUPRC
LungDWM-w/oAtns	0.918±0.012•	0.867±0.063•	0.912±0.011 •	0.856±0.077•
LungDWM-w/oGANs	0.920±0.009•	0.940±0.011•	0.915±0.010•	0.939±0.010•
LungDWM-GANs w/oAtns	0.930±0.011•	0.948±0.013	0.923±0.009•	0.948±0.012
LungDWM-w/oSLoss	0.922±0.019•	0.947±0.017	0.918±0.021•	0.941±0.019•
LungDWM-w/oILoss	0.927±0.018•	0.945±0.019•	0.920±0.020•	0.940±0.021•
LungDWM	0.942±0.011	0.961±0.011	0.937±0.012	0.958±0.014

**Fig. 2.** LungDWM under different values of individual factor λ

To further study whether LungDWM can effectively integrate multiomics data, we record the diagnostic performance of the original single-omics data under the same sample pool. Particularly, missing omics samples are filled with average values of available samples in the same omics. In addition, we use GAN-imputed omics data for diagnosis and report the results in Table 4. We can observe that: (i) Compared with diagnostic results from single omics data filled with average values, the GAN imputed omics data all enable a better diagnosis performance and the most prominent improvement is the image omics. This suggests that the omics data imputed by our GAN module has the similar distributions as the available data and can be used to enrich patient features, thereby improving the diagnostic performance. (ii) LungDWM can effectively integrate multiomics data for disease diagnosis. Compared to the diagnosis using single omics data, LungDWM achieves a very significant performance improvement. This proves that the fusion of multiomics data makes more diverse patient features, which reduces the interference of noise and low coverage of single-omics data and gives more accurate diagnosis.

For a more comprehensive study, we performed more experiments on LungDWM by excluding one particular omics data and multiomics data with different missing rates. We observe that LungDWM can achieve more accurate diagnosis by effectively integrating more types of omics data, and excluding transcriptomics or image data has a serious performance drop than other omics data. We also observe that LungDWM maintains a better robustness on weakly paired multiomics data with different missing rates. The results and analysis are provided in Supplementary Section S3 of Supplementary file.

3.3 Model interpretation

To verify the authenticity of LungDWM for pathologists, we study the attention weights ω^v obtained by LungDWM on genomics, epigenetics and transcriptomics data, these weights also signify the importance of selected features for diagnosis. We separately sum the attention weights of each omics data and average them under 10-fold cross-validation, then select several feature sites with larger weights for the following analysis and validation. For the genomics data, we investigate 14 genes with the highest weights from 19 728 gene mutation sites. For the epigenetics data, we study 14 sites with the highest weights among 9816 selected methylated CpG sites. As

for transcriptomics data, we select 16 miRNAs with the highest weights among 1881 miRNAs. Some highly weighted molecules are revealed in Figure 3a–c.

In the previous experiments, we proved that the attention weights can improve the diagnosis performance by picking out important features from each omics data. For genomics data, we initially assigned the same weight ($5.068e-5$) to 19 728 genes sites. After the model is optimized, the attention weights (after Softmax normalization) under 10-fold cross-validation are averaged, and the highest weight is 9.22 times larger than the initial one. This pattern also applies to the attention weights of epigenetics and transcriptomics data. The initial attention weights for 9816 CpG sites and 1881 miRNAs are $1.018e-4$ and $5.316e-4$. After optimization, they have increased by 52% and 17%, respectively, than the initial one. This fact shows that LungDWM can pick out important features for diagnosing cancer subtypes and can aid pathologists to identify causal and therapeutic sites via the attention weights learned from multiomics data. More discussions and biomedical evidences of these sites are given in Supplementary Section S4 of Supplementary file.

We find a large number of highlighted sites have been verified to be (potentially) associated with LUAD, LUSC and lung cancer. We also find out some novel feature sites associated with lung cancer. These prove the authenticity of LungDWM in subtype diagnosis and its potential to assist pathologists for targeted therapy. In addition, we perform statistics on these sites of genomics and epigenetics group, and observe patients of different subtypes with varying mutation propensities at these sites. These statistical results are consistent with previous biological experiments (Herbst *et al.*, 2018; Vargas and Harris, 2016), and also justify the rationality of extracting shared and specific features to preserve the variety of subtypes.

A miRNA can regulate the expression of multiple genes, and the expression of a gene can also be regulated by multiple miRNAs. We also analyze the gene sites captured by the genomics and epigenetics group, and miRNAs captured by the expression group, we study the biological targets predicted by these miRNA targets with 24 gene sites in the genome and epigenetics group from TargetScan database (McGeary *et al.*, 2019), and create a heatmap in Supplementary Figure S4 of Supplementary file. We find the identified miRNA targets are mostly associated with genes with mutation sites and CpG sites. This not only further confirms the authenticity of the identified important miRNAs by attention weights for lung cancer subtype diagnosis but also proves that LungDWM can effectively integrate correlated features of multiomics data, so as to make accurate subtype diagnosis. More results are given in Supplementary Section S4.3 of Supplementary file.

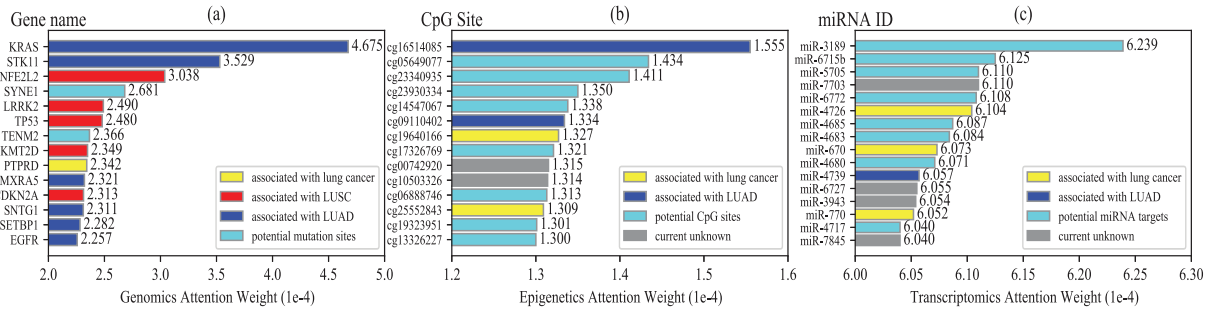
These results not only verify the authenticity and interpretability of attention weights in multiomics data but also suggest the application values of LungDWM for clinical diagnosis and targeted therapy.

4 Conclusion

In this article, we propose LungDWM for lung cancer subtype diagnosis using weakly paired multiomics data. LungDWM leverages attention-based feature encoders to extract the shared and specific features of multiomics data, imputes missing omics data from available ones through generative adversarial learning and makes the subtype diagnosis by fusing real and imputed data. Experimental

Table 4. Results of LungDWM using single-/multiomics data

Variant	Accuracy	AUROC	F1-Score	AUPRC
Only Genomics	0.812±0.019	0.896±0.015	0.809±0.023	0.901±0.021
Imputed Genomics	0.827±0.008	0.913±0.010	0.829±0.009	0.918±0.011
Only Epigenetics	0.826±0.014	0.915±0.011	0.799±0.014	0.920±0.011
Imputed Epigenetics	0.841±0.013	0.930±0.013	0.817±0.015	0.937±0.012
Only Transcriptomics	0.829±0.015	0.919±0.017	0.823±0.013	0.910±0.019
Imputed Transcriptomics	0.845±0.013	0.927±0.015	0.847±0.011	0.923±0.015
Only Images	0.805±0.019	0.892±0.016	0.798±0.021	0.886±0.023
Imputed Images	0.838±0.014	0.922±0.012	0.837±0.014	0.925±0.009
LungDWM	0.942±0.011	0.961±0.011	0.937±0.012	0.958±0.014

**Fig. 3.** Top-ranked Gene names, CpG sites and miRNA targets and corresponding attention weights obtained from each omics data encoder

results show that LungDWM not only can more accurately diagnose cancer subtypes than state-of-the-art methods but also enable a high authenticity and good interpretability. We will expand LungDWM for other cancers and to predict their prognosis in federated learning framework to protect the privacy of multiomics data.

Funding

This work is supported by the National Natural Science Foundation of China (61872300, 62072380 and 62272276) and Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) (NO.2021CXGC010506).

Conflict of Interest: The authors declare that there is no conflict of interest.

References

Ahmed, K.T. et al. (2021) Multi-omics data integration by generative adversarial network. *Bioinformatics*, 38, 179–186.

Arjovsky, M. et al. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223.

Arya, N. and Saha, S. (2021) Generative incomplete multi-view prognosis predictor for breast cancer: GIMPP. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 99, 1–13.

Cheerla, A. and Gevaert, O. (2019) Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*, 35, i446–i454.

Coudray, N. et al. (2018) Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.*, 24, 1559–1567.

Crowley, E. et al. (2013) Liquid biopsy: monitoring cancer-genetics in the blood. *Nat. Rev. Clin. Oncol.*, 10, 472–484.

Gligorijević, V. and Pržulj, N. (2015) Methods for biological data integration: perspectives and challenges. *J. R. Soc. Interface*, 12, 20150571.

Goodfellow, I. et al. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680.

Gulrajani, I. et al. (2017). Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pp. 5769–5779.

Hao, X. et al. (2017) DNA methylation markers for diagnosis and prognosis of common cancers. *Proc. Natl. Acad. Sci. U S A*, 114, 7414–7419.

Herbst, R.S. et al. (2018) The biology and management of non-small cell lung cancer. *Nature*, 553, 446–454.

Howlader, N. et al. (2020) The effect of advances in lung-cancer treatment on population mortality. *N Engl. J. Med.*, 383, 640–649.

Lehman, C.D. and Wu, S. (2021) Stargazing through the lens of AI in clinical oncology. *Nat. Cancer*, 2, 1265–1267.

Li, Y. et al. (2018) A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.*, 19, 325–340.

McGeary, S.E. et al. (2019) The biochemical basis of microRNA targeting efficiency. *Science*, 366, eaav1741.

Menyhárt, O. and Györfy, B. (2021) Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Comput. Struct. Biotechnol. J.*, 19, 949–960.

Mobadersany, P. et al. (2018) Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. U S A*, 115, E2970–E2979.

Rappoport, N. and Shamir, R. (2019) NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, 35, 3348–3356.

Su, Y. et al.; ISB-Swedish COVID19 Biobanking Unit. (2020) Multi-omics resolves a sharp disease-state shift between mild and moderate covid-19. *Cell*, 183, 1479–1495.e20.

Sun, D. et al. (2019) A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 16, 841–850.

Tan, Q. et al. (2021) Individuality- and commonality-based multiview multilabel learning. *IEEE Trans. Cybern.*, 51, 1716–1727.

Vargas, A.J. and Harris, C.C. (2016) Biomarker development in the precision medicine era: lung cancer as a case study. *Nat. Rev. Cancer*, 16, 525–537.

Vincent, P. et al. (2010) Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11, 3371–3408.

Wang, Q. et al. (2020). Multimodal learning with incomplete modalities by knowledge distillation. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1828–1838.

Wang, Q. et al. (2021a) Generative partial multi-view clustering with adaptive fusion and cycle consistency. *IEEE Trans. Image Process.*, 30, 1771–1783.

- Wang,X. *et al.* (2021b) Lung cancer subtype diagnosis by fusing image-genomics data and hybrid deep networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **99**, 1–12.
- Weinstein,J.N. *et al.*; Cancer Genome Atlas Research Network. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
- Yang,H. *et al.* (2021) Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data. *Bioinformatics*, **37**, 2231–2237.
- Yuan,L. *et al.* (2012). Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1149–1157.
- Zappa,C. and Mousa,S.A. (2016) Non-small cell lung cancer: current treatment and future advances. *Transl. Lung Cancer Res.*, **5**, 288–300.
- Zhang,C. *et al.* (2022) Deep partial multi-view learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, **44**, 2402–2415.