Yao *et al. BioData Mining* (2024) 17:9
https://doi.org/10.1186/s13040-024-00360-6

BioData Mining

RESEARCH Open Access

MOCAT: multi-omics integration with auxiliary classifiers enhanced autoencoder

Xiaohui Yao[1,2], Xiaohan Jiang[1], Haoran Luo[1,2], Hong Liang[2], Xiufen Ye[2], Yanhui Wei[2] and Shan Cong[1,2*]

*Correspondence:
Shan.Cong@hrbeu.edu.cn

[1] Qingdao Innovation and Development Center, Harbin Engineering University, 1777 Sansha Rd, Qingdao 266000, Shandong, China
[2] College of Intelligent Systems Science and Engineering, Harbin Engineering University, 145 Nantong St, Harbin 150001, Heilongjiang, China

Abstract

Background: Integrating multi-omics data is emerging as a critical approach in enhancing our understanding of complex diseases. Innovative computational methods capable of managing high-dimensional and heterogeneous datasets are required to unlock the full potential of such rich and diverse data.

Methods: We propose a Multi-Omics integration framework with auxiliary Classifiers-enhanced AuToencoders (MOCAT) to utilize intra- and inter-omics information comprehensively. Additionally, attention mechanisms with confidence learning are incorporated for enhanced feature representation and trustworthy prediction.

Results: Extensive experiments were conducted on four benchmark datasets to evaluate the effectiveness of our proposed model, including BRCA, ROSMAP, LGG, and KIPAN. Our model significantly improved most evaluation measurements and consistently surpassed the state-of-the-art methods. Ablation studies showed that the auxiliary classifiers significantly boosted classification accuracy in the ROSMAP and LGG datasets. Moreover, the attention mechanisms and confidence evaluation block contributed to improvements in the predictive accuracy and generalizability of our model.

Conclusions: The proposed framework exhibits superior performance in disease classification and biomarker discovery, establishing itself as a robust and versatile tool for analyzing multi-layer biological data. This study highlights the significance of elaborated designed deep learning methodologies in dissecting complex disease phenotypes and improving the accuracy of disease predictions.

Keywords: Multi-omics integration, Auxiliary classifier, Autoencoder, Attention mechanism, Trustworthy learning, Disease prediction

Introduction

Recent advancements in omics technologies have enabled large-scale data acquisition across multiple biological layers, including genomics, transcriptomics, proteomics, metabolomics, and many others. Considering that each type of omics data contributes distinct layers of biological information, data integration serves as an efficient tool in multi-omics studies, for not only providing a comprehensive understanding of the multifaceted complexity inherent in biological phenomena but also substantially improves our capabilities in elucidating disease mechanisms and identifying disease biomarkers [1, 2].

Compared to single-omics studies, multi-omics analyses require computational methodologies to encompass a more holistic perspective. These advanced methods aim to overcome the inherent limitations of single-omics approaches by providing nuanced insights into complex biological interactions. However, the complexity and heterogeneity contained within and across multi-omics data pose significant challenges to their integration and downstream tasks. Traditional approaches for analyzing multi-omics data mainly rely on statistical methodologies, as indicated in studies [3–5]. These statistical approaches often encounter difficulties in feature extraction, typically requiring manual intervention that can be both labor-intensive and suboptimal for interpreting global feature significance. Deep learning models have been widely designed to address this issue and demonstrated superior predictive capabilities and proficiency in identifying nonlinear and hierarchical features [6–8].

Deep neural networks in multi-omics analyses enable the autonomous extraction of relevant features and facilitate the identification of intricate associations among them. However, when applying neural networks to multi-omics data, a series of challenges remain, one notable issue being the 'curse of dimensionality' [9]. Due to the multiple causes and pathogenic mechanisms underlying complex diseases, omics data are particularly prone to this issue. Such high dimensionality creates intricate spatial distributions that can impede both traditional machine learning and deep learning algorithms in their classification tasks. Preserving all original features magnifies the computational complexity in such high-dimensional spaces, inducing overfitting and diminishing the predictive accuracy. To address these challenges, autoencoder models have been investigated as a means to transform and integrate multi-omics features, particularly for discerning disease subtypes [10–12]. Various strategies based on autoencoders have been proposed to integrate high-dimensional, multi-source datasets and to derive low-dimensional latent representations. For example, Wang et al. [13] employed autoencoders to align and integrate data from single-cell RNA-seq and ATAC-seq, adeptly mapping the sparse and noisy data from varied spaces into a harmonized subspace for improved alignment and integration. Lin et al. [14] introduced scMDC, an architecture featuring one encoder for cascading data and two decoders for each data modality. This design uniquely characterizes distinct data sources and co-learns deep embeddings of latent features for cluster analyses. Autoencoders use a combination of nonlinear functions to reconstruct the original inputs, which can be used as new feature representations of the original data. These algorithms have been proven effective in producing clinically relevant features [15], analyzing high-dimensional omics data [16, 17], and integrating multi-omics data [7, 18]. However, autoencoders can exhibit suboptimal performance in certain tasks, especially when the generated subrepresentations are utilized for downstream tasks [7, 19]. This issue often originates from the focus of traditional autoencoder objective functions on input reconstruction, which can limit their effectiveness in classification tasks.

In addition to omics-specific feature extraction, data fusion is another key step in multi-omics studies. Attention mechanisms have been a robust technique for enhancing classification performance by selectively emphasizing salient features across various omics levels. These mechanisms enhance the model performance by prioritizing more relevant features for classification outcomes. Such adaptability is particularly

crucial in light of the varying importance of different features. As substantiated by feature interpretation studies such as Mogonet [20], attention mechanisms are justified in their application for multi-omics fusion, given their capacity to weigh the importance of features in diagnosis prediction adaptively.

On the other hand, deep learning models for classification tasks typically adopt maximum class probability (MCP), that is, the highest probability values given by the softmax output, to evaluate the prediction confidences [21]. This can lead to assigning high confidence values to even incorrect predictions. To mitigate this limitation and enhance classification accuracy, the true class probability (TCP) criterion is integrated into the loss function [22, 23]. Unlike MCP, TCP assesses the predicted probability for each class against the probability of the true class label, incorporating these values into the loss computation. This criterion acts as a regularizer during training by offering more detailed insights into the performance of the classifier on individual sample predictions. This becomes more essential in challenging scenarios, like those where performance improvement is hindered due to hard samples (that is commonly observed in complex diseases) or when only a few samples are incorrectly predicted due to the well-designed models.

Upon recognizing the observed limitations in existing methods, we present a Multi-Omics integration framework with auxiliary Classifiers-enhanced AuToencoder (MOCAT) to improve both stability and predictive accuracy in disease classification tasks. Acknowledging the importance of explicability within the domain of biomedical research, our framework also incorporates model interpretability for biomarker discovery. The architecture employs autoencoders for efficient high-dimensional feature compression, while the integration of omics-specific classifiers promotes refined optimization aligned with disease prediction. Furthermore, adopting attention mechanisms affords greater flexibility in fusing multiple omics types. We also incorporate the trustworthy strategy to facilitate fine-grained optimization in the weighting of the classification network, culminating in an augmented accuracy of classification. Benchmark experiments and comparative evaluations show that the proposed model outperforms existing state-of-the-art methods, with extensive validations demonstrating both the reliability and interpretability of the proposed framework.

Our main contributions are summarized as the following:

- **Omics-Specific Feature Optimization:** We introduce auxiliary classifiers tailored for each type of omics data, which significantly enhances feature representation by identifying the most informative biomarkers pertinent to disease states.
- **Enhanced Classifier Confidence Calibration:** We incorporate the true class probability criterion to regularize classifier confidence of incorrect predictions, thereby improving model overconfidence and enhancing predictive accuracy.
- **Explainability:** By integrating mechanisms that elucidate the decision-making process of our model, we provide predictive proficiency and facilitate a deeper understanding of the underlying biological phenomena, thereby aiding in the interpretive aspects of biomarker discovery.

- **State-Of-The-Art (SOTA) Performance:** The proposed framework has yielded superior results on four independent datasets, indicating an improvement over the current benchmarks.

Materials and methods

Datasets

To conduct fair comparative experiments, we adopt public data preprocessed by Wang et al. [20], which provide four datasets for multi-omics disease classification, including a binary classification ROSMAP dataset for Alzheimer's disease (AD) patients and normal controls (NC), a BRCA dataset for PAM50 subtype classification of invasive breast cancer (five-class), an LGG dataset for the grade classification of gliomas (binary), and a KIPAN dataset for subtype classification of renal cancer (three-class). Table 1 shows the detailed information of these datasets, where the preprocessed features were used for training.

Model formulation

The proposed model is structured into three sequential phases for comprehensive data analysis. Phase 1 focuses on efficiently compressing high-dimensional data to extract critical omics-specific features. Phase 2 involves the integration of multi-omics data and confidence-based disease prediction. Finally, Phase 3 is concerned with biomarker discovery, harnessing the insights collected from the previous phases. The overall architecture of the proposed model is shown in Fig. 1.

Phase 1: omics-specific feature extraction

In phase 1, high-dimensional features of each omics dataset are fed into autoencoder networks for extracting representative features. At the same time, each omics data is separately trained to assist the autoencoder network in learning a more compact and accurate representation. Developing independent models for each omics dataset can help avoid losing the specificity of each data source, as they may exhibit distinct dynamics. The independent models are expected to provide reliable feature information for multimodal fusion.

Autoencoders for dimensionality reduction: Three autoencoders are trained separately on different omics types. In particular, each autoencoder uses a multi-objective optimization method in the encoder with Dropout, BatchNorm, and ELU

Table 1 Summary of datasets

Dataset	Sample	Number of raw features			Number of features for training		
		mRNA	methy	miRNA	mRNA	methy	miRNA
ROSMAP	NC: 169, AD: 182	55,889	23,788	309	200	200	200
BRCA	Luminal A: 436, Luminal B: 147, HER2-enriched: 46, Normal-like: 115, Basal-like: 131	20,531	20,106	503	1,000	1,000	503
LGG	Grade 2: 246, Grade 3: 264	20,531	20,114	548	2,000	2,000	548
KIPAN	KICH: 66, KIRC: 318, KIRP: 274	20,531	20,111	445	2,000	2,000	445

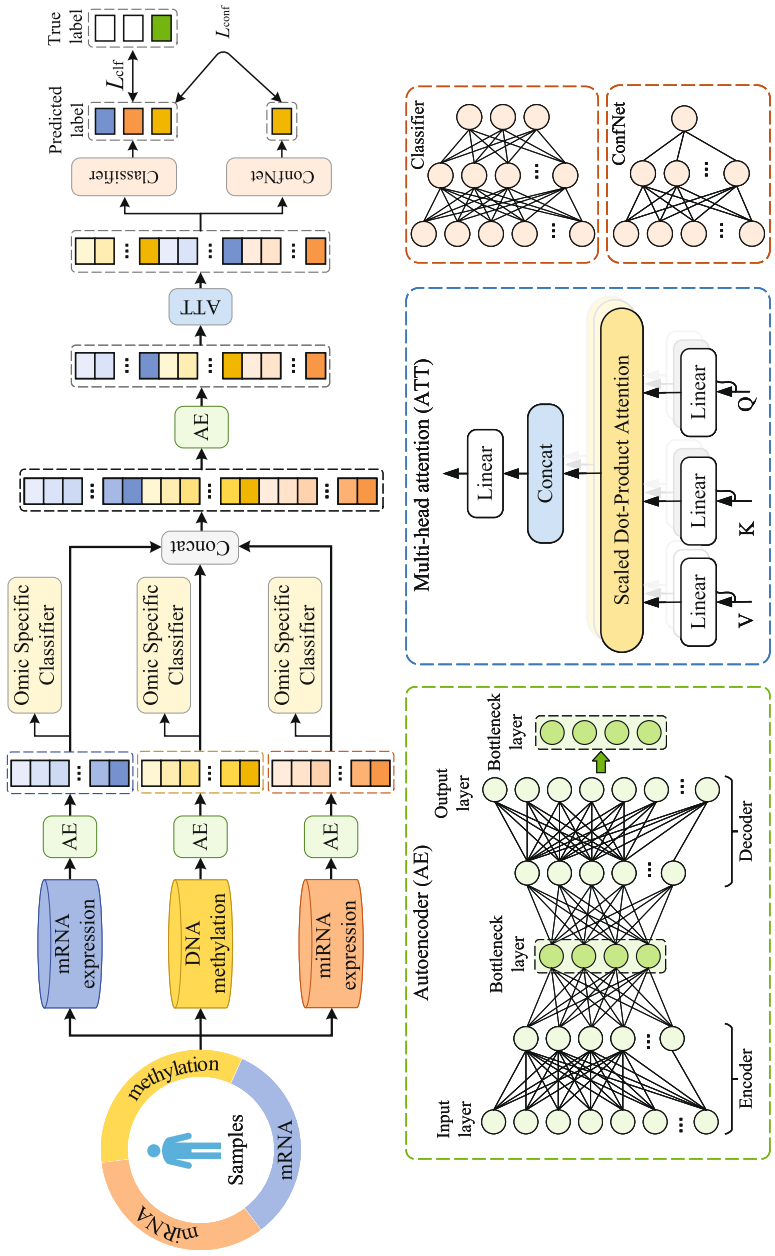

Fig. 1 Framework of the MOCAT. The top panel shows the overall architecture of the proposed model: 1) high-dimensional features of multiple omics datasets are fed into an autoencoder network for dimensionality reduction to obtain representative features; 2) three omics-specific auxiliary classifiers are trained to assist in learning more compact and accurate sub-representations; 3) feature sub-representations are fused and further compressed by an autoencoder network a self-attention module to fuse the complementary information embedded across multi-omics adaptively; 4) confident network (ConfNet) is employed to adjust the prediction confidence linked to the fused features adaptively. The bottom panel illustrates the detailed architectures of the autoencoder, attention, and confidence networks

activation functions to compress the original data and extract corresponding low-dimensional representations. The Dropout layer is generally set after the fully connected layer and randomly drops part of the nodes according to a preset ratio during training, effectively improving the problem of model overfitting and improving model generalization. The BatchNorm layer, as shown in Eq. 1, is used to normalize the mean and variance of features and is widely applied in deep learning tasks. This approach speeds up model training and improves model stability while alleviating issues like gradient vanishing and gradient explosion.

$$\text{BatchNorm}(\mathbf{x}) = \gamma \frac{\mathbf{x}_i - \mu_{\mathbf{x}}}{\sqrt{\sigma_{\mathbf{x}}^2 + \epsilon}} + \beta, \tag{1}$$

where $\mu_{\mathbf{x}}$ and $\sigma_{\mathbf{x}}^2$ are the batch mean and variance, γ is a scale parameter used to scale the normalized data, ϵ is a small constant added to the denominator to prevent division by zero, and β is a shifting parameter used to shift the normalized results by the batch mean.

The ELU activation function can better manage the gradient vanishing problem, making the training converge faster, thus achieving better results:

$$\text{ELU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha(e^x - 1) & \text{if } x < 0 \end{cases}. \tag{2}$$

In detail, each autoencoder model consists of multiple fully connected layers, with a bottleneck layer in the middle that minimizes node size. The intermediate layer with the fewest nodes serves as the bottleneck layer for dimensionality reduction in the original dataset. The last layer reconstructs the raw data from the first layer. We aim to minimize reconstruction errors and extract better feature representations at the bottleneck layer. The effectiveness of feature representation can be evaluated by omics-specific classifiers.

Omics-specific classifiers: Auxiliary classifiers specific to each omics data are incorporated to fulfill two objectives: (i) improve representation by encouraging the autoencoders to learn more fine-grained and discriminative features; (ii) improve prediction performance by forcing the model to learn from multiple perspectives. The classification loss from the omics-specific classifiers directs the autoencoders in refining feature compression, ensuring that the compressed representations align with the distinguishing characteristics of each omics data. This alignment is intuitively thought to promote the overall effectiveness of the autoencoder for dimensionality reduction.

Overall, M omics-specific feature subrepresentations $\mathrm{F}^{(m)}, m \in \{1, \ldots, M\}$ were obtained from phase 1. The loss of the first phase includes the reconstruction loss $\mathcal{L}_{\text{rc}}^{(m)}$ of each autoencoder and the auxiliary classification loss $\mathcal{L}_{\text{ac}}^{(m)}$ of each omics-specific classifier, and can be expressed as:

$$\mathcal{L}_{\text{phase1}} = \lambda_1 \sum_{m=1}^{M} \mathcal{L}_{\text{rc}}^{(m)} + \lambda_2 \sum_{m=1}^{M} \mathcal{L}_{\text{ac}}^{(m)}, \tag{3}$$

where λ_1 and λ_2 are hyperparameters for adjusting different losses. We set $\lambda_1 = 1$ and $\lambda_2 = 0.005$ in our experiment.

Phase 2: cross-omics fusion and trustworthy prediction

Integrating heterogeneous multi-omics data, characterized by varying expression patterns and dimensions, presents a significant challenge. In phase 2 of our approach, we delve into improving both multi-omics fusion and final disease prediction, utilizing feature representations derived from phase 1. We specifically apply the attention mechanism to establish global correlations within the fused features and introduce the classification confidence mechanism into the network to enhance its prediction performance.

Adaptive fusion with autoencoder and attention: We first concatenated the omics-specific subrepresentations obtained from phase 1 to create the preliminary fused representations. Subsequently, an autoencoder network was trained to map these heterogeneous features into a novel embedding space. This space is designed to learn and encapsulate shared representations across the different omics data types. This procedure facilitates the extraction of discriminative and representative features from a more comprehensive perspective, thereby augmenting the overall effectiveness of the model. Given the M omics-specific feature representations $\mathrm{F}^{(m)}, m \in \{1, \ldots, M\}$, the output $\mathbf{Z}_{\mathrm{AE}}$ of the autoencoder layer is calculated as follows:

$$\mathbf{Z}_{\mathrm{AE}} = \mathrm{Autoencoder}\left(\mathop{\Big\Vert}_{m=1}^{M} \mathrm{F}^{(m)} \right), \tag{4}$$

where $\|$ represents concatenation operation.

The integrated features were subsequently fed into an attention layer. This is designed to capture and emphasize the distinct significance of various omics modalities, thereby augmenting the efficacy of our model. It has been noted that different types of omics data contribute variably to the aggregate predictive accuracy. The integration of the attention mechanism allows for a dynamic recalibration of the influence exerted by the fused modality features during the classification procedure. Given the input $\mathbf{Z}_{\mathrm{AE}}$, the output $\mathbf{Z}_{\mathrm{Att}}$ of the attention layer is calculated as follows:

$$\mathbf{Z}_{\mathrm{Att}} = \mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Attention}\left(\mathbf{Z}_{\mathrm{AE}}\mathbf{W}^{Q}, \mathbf{Z}_{\mathrm{AE}}\mathbf{W}^{K}, \mathbf{Z}_{\mathrm{AE}}\mathbf{W}^{V}\right), \tag{5}$$

where $\mathbf{W}^{Q} \in \mathbb{R}^{d \times d_q}$, $\mathbf{W}^{K} \in \mathbb{R}^{d \times d_k}$, and $\mathbf{W}^{V} \in \mathbb{R}^{d \times d_v}$ with $d_q = d_k$ are three learnable weight matrices for generating the corresponding matrices of query $\mathbf{Q} \in \mathbb{R}^{n \times d_q}$, key $\mathbf{K} \in \mathbb{R}^{n \times d_k}$, and value $\mathbf{V} \in \mathbb{R}^{n \times d_v}$, n is the number of samples and d is the embedding dimensionality of the previous autoencoder layer.

Trustworthy prediction with ConfNet: In addition to improving the representation effectiveness of intra- and inter-omics data, we also employed the trustworthy strategy to assess and adaptively adjust the prediction confidence linked to the fused features.

The traditional method for determining confidence in classification, known as the maximum class probability (MCP), relies on the highest probability output of the softmax function. For a given input feature matrix $\mathbf{Z}_{\mathrm{Att}}$, the classifier acts as a probabilistic model. This model assigns a predictive probability distribution $P(Y|\mathbf{Z}_{\mathrm{Att}})$ for each class in the set $k = \{1, ..., K\}$. The class with the highest probability is then selected as the

predicted class, denoted as $\hat{y} = \arg\max_{k\in\{1,\ldots,K\}} P(Y=k|\mathbf{Z}_{\text{Att}})$. However, a notable issue with MCP is its tendency to exhibit overconfidence for incorrect predictions.

The true class probability (TCP) confidence criterion was proposed to solve this problem [22], by assigned confidences according to $P(Y=\mathbf{y}^*|\mathbf{Z}_{\text{Att}})$, where $\mathbf{y}^*$ represents the true label vector. TCP and MCP yield equivalent results when a sample is correctly classified. However, for misclassified samples, TCP provides a more conservative and, thus, potentially more accurate confidence value. The direct estimation of TCP confidence on the test set is not feasible due to the absence of true labels. To solve this problem, a confidence network (denoted as ConfNet) was introduced to the training data, and the parameters were learned as follows:

$$\mathcal{L}_{\text{Conf}} = \left\| \text{Conf}(\mathbf{Z}_{\text{Att}}, \mathbf{y}^*) - \widehat{\text{Conf}}(\mathbf{Z}_{\text{Att}}, \text{ConfNet}(\cdot)) \right\|^2, \tag{6}$$

$$\text{Conf}(\mathbf{Z}_{\text{Att}}, \mathbf{y}^*) = f(\mathbf{Z}_{\text{Att}}, \mathbf{y}^*), \tag{7}$$

$$\widehat{\text{Conf}}(\mathbf{Z}_{\text{Att}}, \text{ConfNet}(\cdot)) = f(\mathbf{Z}_{\text{Att}}, \text{ConfNet}(\cdot)). \tag{8}$$

Here, the function $\text{Conf}(\mathbf{Z}_{\text{Att}}, \mathbf{y}^*)$, denoted by f, processes the output $\mathbf{Z}_{\text{Att}}$ from the attention layer to produce the true label vector $\mathbf{y}^*$ through a fully connected layer. The $\widehat{\text{Conf}}(\mathbf{Z}_{\text{Att}}, \text{ConfNet}(\cdot))$ follows the same logic. In this context, $\text{Conf}(\mathbf{Z}_{\text{Att}}, \mathbf{y}^*)$ is the TCP confidence proposed to learn. As illustrated in Fig. 1, both the confidence network and the classifier were built upon the output of the attention layer. The classifier was trained using the cross-entropy loss and fixed, after which the confidence network was trained according to Eq. 6. In this way, the model is designed to adjust the feature weights in response to misclassified samples adaptively. This dynamic penalization mechanism enhances the model to learn from errors and refine its predictive accuracy. Furthermore, by effectively distinguishing false predictions from true ones through enhanced confidence separation, the TCP criterion holds the potential to boost the generalizability of the model and thereby reduce the risk of overfitting.

Therefore, the loss of phase 2 consists of the reconstruction loss of the fused features $\mathcal{L}_{\text{rc}}$ and the confidence loss $\mathcal{L}_{\text{conf}}$:

$$\mathcal{L}_{\text{phase2}} = \lambda_3 \mathcal{L}_{\text{rc}} + \lambda_4 \mathcal{L}_{\text{conf}}, \tag{9}$$

where λ_3 and λ_4 are hyperparameters used to balance different losses and are set to 0.5 in the experiments.

In total, the loss of the entire model includes the phase 1 loss, phase 2 loss, and the cross-entropy loss for the final classification $\mathcal{L}_{\text{clf}}$:

$$\mathcal{L} = \mathcal{L}_{\text{phase1}} + \mathcal{L}_{\text{phase2}} + \mathcal{L}_{\text{clf}}, \tag{10}$$

Phase 3: biomarkers identification

Identifying biomarkers is fundamental for understanding underlying biological mechanisms and interpreting outcomes in biomedical contexts. Discovering biomarkers via deep learning models facilitates the discernment of highly representative and predictive

features in classification tasks. We evaluated the importance of each omics feature to find those showing significant effects on the prediction performance of the model. Specifically, feature ablation was employed wherein each feature was eliminated, and the feature-level importance score was calculated according to the decreasing accuracy. In practice, we repeated five times to obtain the mean importance measurements to reduce experimental variability.

Furthermore, we investigated the inter-omics relevance of the identified biomarkers to showcase the efficacy of the proposed model in uncovering interactive cross-omics biomarkers. We specifically selected the top 30 biomarkers identified through the previous feature ablation analysis. Subsequently, we evaluated the joint effects of all possible combinations of inter-omics biomarkers, including both pairwise (e.g., mRNA-methylation, mRNA-miRNA, and methylation-miRNA) and tri-omics (mRNA-methylation-miRNA) interactions. Moreover, we randomly selected 1, 000 sets of tri-omics biomarker combinations and compared their prediction importance with our top findings to demonstrate the significant inter-omics relevance of the biomarkers prompted by our method.

Results

We evaluated the performance of our proposed method by comparing it with state-of-the-art multi-omics classification approaches using four public datasets. Furthermore, extensive ablation studies were executed to elucidate the efficacy of each component within our framework. We focused on three metrics for binary classification: classification accuracy (ACC), F1 score, and area under the ROC curve (AUC). For multiclass classification datasets, we also focused on three metrics including accuracy (ACC), weighted average F1 score (F1_w), and macroaverage F1 score (F1_m).

Diseases prediction comparison

Our comparative analysis encompassed fourteen computational methods, including six early-stage single-omics benchmark algorithms, namely, K-nearest neighbors (KNN) [24], support vector machine (SVM) [25], Lasso [26], random forest (RF) [27], eXtreme Gradient Boosting (XGboost) [28], and fully connected neural networks (NN) [29]. We also evaluated seven advanced multi-omics classification frameworks, which include group-regularized ridge regression (GRridge) [30], Bayesian partial least squares discriminant analysis-based BPLSDA [31], BSPLSDA [31], Concatenate Fusion (CF) for post-modality connection of multi-omics representations [32], Gate Modulated Unit (GMU) for information fusion with gating mechanisms [33], and the two state-of-the-art algorithms Mogonet [20] and Dynamics [34]. For fair comparisons, we evaluated all methods according to the same experimental settings as Mogonet [20], and the outcomes were expressed as the mean and 95% confidence interval (95% CI) of five experiments.

As shown in Tables 2 and 3, our model outperformed both the benchmark and state-of-the-art methods on both binary and multiclass classification tasks. Our approach consistently outperformed existing methods on the ROSMAP, BRCA, and LGG datasets, demonstrating the robustness and adaptability of our model. In the case of the KIPAN dataset, performance from our model was on par with advanced algorithms, validating its competitive capability. Statistical analysis demonstrated that the performance

Table 2 Comparison with state-of-the-art methods on ROSMAP and BRCA datasets

Method	ROSMAP (2 Categories)			BRCA (5 Categories)		
	ACC(%) (95% CI)	F1(%) (95% CI)	AUC(%) (95% CI)	ACC(%) (95% CI)	F1_w(%) (95% CI)	F1_m(%) (95% CI)
KNN	65.7 (61.2-70.2)	67.1 (61.6-72.6)	70.9 (65.3-76.5)	74.2 (71.2-77.2)	73.0 (70.1-75.9)	68.2 (65.1-71.3)
SVM	77.0 (74.0-80.0)	77.8 (75.8-79.8)	77.0 (73.8-80.2)	72.9 (70.7-75.1)	70.2 (68.3-72.1)	64.0 (61.9-66.1)
Lasso	69.4 (64.8-74.0)	73.0 (68.9-77.1)	77.0 (72.7-81.3)	73.2 (71.7-74.7)	69.8 (67.9-71.7)	64.2 (61.0-67.4)
RF	72.6 (69.0-76.2)	73.4 (70.8-76.0)	81.1 (78.7-83.5)	75.4 (74.3-76.5)	73.3 (72.1-74.5)	64.9 (63.3-66.5)
XGBoost	76.0 (70.3-81.7)	77.2 (71.6-82.8)	83.7 (80.0-87.4)	78.1 (77.1-79.1)	76.4 (75.2-77.6)	70.1 (68.0-72.2)
NN	75.5 (72.9-78.1)	76.4 (73.8-79.0)	82.7 (79.6-85.8)	75.4 (71.9-78.9)	74.0 (69.8-78.2)	66.8 (61.0-72.6)
GRridge	76.0 (71.8-80.2)	76.9 (73.3-80.5)	84.1 (81.2-87.0)	74.5 (72.5-76.5)	72.6 (70.2-75.0)	65.6 (62.5-68.7)
BPLSDA	74.2 (71.2-77.2)	75.5 (72.6-78.4)	83.0 (79.9-86.1)	64.2 (63.1-65.3)	53.4 (51.7-55.1)	36.9 (34.8-39.0)
BSPLSDA	75.3 (71.2-79.4)	76.4 (72.1-80.7)	83.8 (81.2-86.4)	63.9 (62.9-64.9)	52.2 (50.2-54.2)	35.1 (32.4-37.8)
CF	78.4 (77.0-79.8)	78.8 (78.2-79.4)	88.0 (87.4-88.6)	81.5 (80.5-82.5)	81.5 (80.4-82.6)	77.1 (76.0-78.2)
GMU	77.6 (74.5-80.7)	78.4 (76.4-80.4)	86.9 (84.9-88.9)	80.0 (75.2-84.8)	79.8 (72.3-86.7)	74.6 (67.4-81.8)
Mogonet	81.5 (78.6-84.4)	82.1 (79.4-84.8)	87.4 (85.9-88.9)	82.9 (80.7-85.1)	82.5 (80.5-84.5)	77.4 (75.3-79.5)
Dynamics	84.2 (83.6-84.8)	84.6 (84.3-84.9)	91.2 (90.9-91.5)	87.7 (87.6-87.8)	88.0 (87.8-88.2)	84.5 (84.3-84.7)
MOCAT(Ours)	**87.6*** (86.7-88.5)	**87.5*** (86.8-88.2)	**92.3*** (91.2-93.4)	**88.5*** (88.1-88.9)	**88.9*** (88.5-89.3)	**86.2*** (85.3-87.1)

Means and 95% confidence intervals (95% CIs) are presented, and the best results are in bold. The 95% CI is calculated using the t-distribution, with degrees of freedom set at $n - 1$, where n is the number of experiments conducted.

Compared to the suboptimal model, the superior model is denoted by * to indicate a statistically significant improvement ($P < 0.05$) when using the two-sample t-test

improvements are significant ($P < 0.05$), further confirming the substantial superiority of the proposed model.

Comparative analysis across varied omics types

In our investigation, we endeavored to integrate diverse omics data types—specifically mRNA, DNA methylation, and miRNA—to provide a more comprehensive understanding of disease etiology and to enhance the accuracy of disease classification beyond what is possible with single or dual omics data sources. This is based on the hypothesis that each data contributes uniquely to the model and that integrating multiple sources can lead to more robust performance. We designed a series of ablative studies on the ROSMAP, BRCA, and LGG datasets to validate our hypothesis, excluding the KIPAN dataset due to its relatively straightforward classification nature. We assessed the performance impact when transitioning from using individual omics datasets to combinations of two and ultimately incorporating all three.

Table 3 Comparison with state-of-the-art methods on LGG and KIPAN datasets

Method	LGG (2 Categories)			KIPAN (3 Categories)		
	ACC(%) (95% CI)	F1(%) (95% CI)	AUC(%) (95% CI)	ACC(%) (95% CI)	F1_w(%) (95% CI)	F1_m(%) (95% CI)
KNN	72.9 (68.7-77.1)	73.8 (69.7-77.9)	79.9 (75.2-84.6)	96.7 (95.3-98.1)	96.7 (95.3-98.1)	96.0 (94.3-97.7)
SVM	75.4 (69.7-81.1)	75.7 (69.5-81.9)	75.4 (69.7-81.1)	99.5 (99.1-99.9)	99.5 (99.1-99.9)	99.4 (98.9-99.9)
Lasso	76.1 (73.9-78.3)	76.7 (74.0-79.4)	82.3 (78.9-85.7)	97.4 (97.2-97.6)	97.4 (97.2-97.6)	97.2 (96.7-97.7)
RF	74.8 (73.3-76.3)	74.2 (73.0-75.4)	82.3 (81.1-83.5)	98.1 (97.4-98.8)	98.1 (97.4-98.8)	97.5 (96.1-98.9)
XGBoost	75.6 (70.6-80.6)	76.7 (72.7-80.7)	84.0 (81.1-86.9)	99.3 (98.3-100)	99.3 (98.3-100)	98.9 (97.2-100)
NN	73.7 (70.8-76.6)	74.8 (71.8-77.8)	81.0 (76.4-85.6)	99.1 (98.5-99.7)	99.1 (98.5-99.7)	99.1 (98.5-99.7)
GRridge	74.6 (69.9-79.3)	75.6 (71.1-80.1)	82.6 (77.1-88.1)	99.4 (98.9-99.9)	99.4 (98.9-99.9)	99.3 (98.8-99.8)
BPLSDA	75.9 (72.8-79.0)	73.8 (70.0-77.6)	82.5 (79.6-85.4)	93.3 (91.7-94.9)	93.3 (91.7-94.9)	91.9 (89.3-94.5)
BSPLSDA	68.5 (65.1-71.9)	66.2 (62.5-69.9)	73.0 (69.8-76.2)	91.9 (90.4-93.4)	91.8 (90.2-93.4)	89.5 (87.8-91.2)
CF	81.1 (79.6-82.6)	82.2 (81.7-82.7)	88.1 (87.6-88.6)	99.9 (99.7-100)	99.9 (99.7-100)	99.9 (99.7-100)
GMU	80.3 (78.4-82.2)	80.8 (79.3-82.3)	88.6 (87.1-90.1)	99.2 (98.6-99.8)	99.2 (98.6-99.8)	98.8 (97.7-99.9)
Mogonet	81.6 (79.6-83.6)	81.4 (79.7-83.1)	84.0 (80.6-87.4)	97.7 (95.7-99.7)	97.6 (95.5-99.7)	95.8 (91.8-99.8)
Dynamics	83.3 (82.8-83.8)	83.7 (83.5-83.9)	88.5 (88.3-88.7)	99.9 (99.8-100)	99.9 (99.8-100)	**99.9** (99.8-100)
MOCAT(Ours)	**85.1*** (84.4-85.8)	**85.1*** (84.1-86.1)	**88.5** (88.0-89.0)	**99.9** (99.8-100)	**99.9** (99.8-100)	99.8 (99.3-100)

Means and 95% confidence intervals (95% CIs) are presented, and the best results are in bold.The 95% CI is calculated using the t-distribution, with degrees of freedom set at $n - 1$, where n is the number of experiments conducted.

Compared to the suboptimal model, the superior model is denoted by * to indicate a statistically significant improvement ($P < 0.05$) when using the two-sample t-test

Figure 2 illustrates the performance comparison of using various omics combinations. It can be observed that utilizing all three omics types yielded the highest performance across all three tasks, except the AUC of mRNA+miRNA on the LGG dataset (89.9% > 88.5%). This emphasizes the advantage of harnessing multiple omics, which provide a more comprehensive spectrum of crucial information. Furthermore, it validates the capacity of our proposed model in effectively extracting and integrating representative features from these diverse omics sources.

Ablation study

We performed ablation studies to assess the key modules used in our method, including the omics-specific auxiliary classifiers (AC), the attention mechanism (Att), and the trustworthy strategy (ConfNet). We respectively removed these three components from the proposed model and explored the prediction performance.

Results are summarized in Table 4. We can observe that each critical component contributes to enhancing the classification efficacy of our model. Specifically, the removal

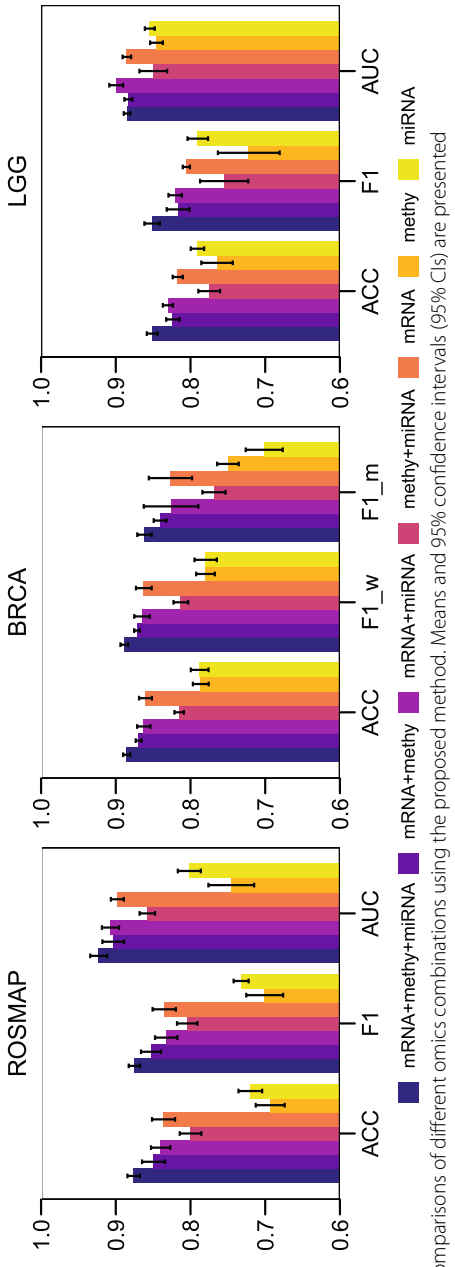

Fig. 2 Performance comparisons of different omics combinations using the proposed method. Means and 95% confidence intervals (95% CIs) are presented

Table 4 Ablation study of the key modules

Dataset	Method	ACC (95% CI)	F1 (95% CI)	AUC (95% CI)
ROSMAP	$\overline{\text{AC}}$: AE_{os}+AE_f+Att+ConfNet	79.4* (76.9-81.9)	79.6* (76.6-82.6)	88.2* (87.3-89.1)
	$\overline{\text{Att}}$: AE_{os}+AC+AE_f+ConfNet	86.7 (85.8-87.6)	86.5 (85.8-87.2)	92.3 (91.8-92.8)
	$\overline{\text{ConfNet}}$: AE_{os}+AC+AE_f+Att	85.5* (84.5-86.5)	85.6* (84.6-86.6)	92.2 (91.3-93.1)
	Ours	**87.6** (86.7-88.5)	**87.5** (86.8-88.2)	**92.3** (91.2-93.4)
LGG	$\overline{\text{AC}}$: AE_{os}+AE_f+Att+ConfNet	80.0* (78.5-81.5)	78.2* (76.5-79.9)	88.8 (88.1-89.5)
	$\overline{\text{Att}}$: AE_{os}+AC+AE_f+ConfNet	83.6* (83.0-84.2)	83.4* (82.5-84.3)	88.9 (88.3-89.5)
	$\overline{\text{ConfNet}}$: AE_{os}+AC+AE_f+Att	84.1* (83.5-84.7)	83.8* (82.8-84.8)	**89.0** (88.5-89.5)
	Ours	**85.1** (84.4-85.8)	**85.1** (84.1-86.1)	88.5 (88.0-89.0)
Dataset	Method	ACC (95% CI)	F1_w (95% CI)	F1_m (95% CI)
BRCA	$\overline{\text{AC}}$: AE_{os}+AE_f+Att+ConfNet	87.8 (87.2-88.4)	88.0 (87.1-88.9)	84.5 (82.4-86.6)
	$\overline{\text{Att}}$: AE_{os}+AC+AE_f+ConfNet	87.4* (86.8-88.0)	87.7* (87.1-88.3)	84.8* (84.7-84.9)
	$\overline{\text{ConfNet}}$: AE_{os}+AC+AE_f+Att	87.9* (87.5-88.3)	88.1* (87.7-88.5)	85.0* (84.4-85.6)
	Ours	**88.5** (88.0-89.0)	**88.9** (88.4-89.4)	**86.2** (85.2-87.2)

Mean values (%) and 95% confidence intervals (CIs) are presented, and the best results are in bold. The 95%CI is calculated using the t-distribution, with degrees of freedom set at $n - 1$, where n is the number of experiments conducted.

The overline denotes the ablation of the corresponding module. The asterisk * denotes a statistically significant difference between the scenarios with and without the respective key module, as computed by the two-sample t-test ($P < 0.05$).

Abbreviations. AC: auxiliary classifiers; AE_{os}: omics-specific autoencoders; AE_f: autoencoders applied on the fused features; Att: self-attention; ConfNet: confidence network

of the expressly designed auxiliary classifiers results in a significant decline in performance. This is particularly noticeable in the context of binary classification tasks. Notably, in the ROSMAP and LGG datasets, we obtain an improvement of 8.2% and 5.1% in accuracy and 7.9% and 6.9% in the F1 score, respectively. This underscores the efficacy of omics-specific classifiers in enriching the capacity of autoencoders for nuanced feature representation.

Incorporating the attention mechanism also enhances almost all of the three experiments. For example, the attention module significantly improves the classification of BRCA subtypes across all three evaluation metrics. This highlights the capacity of the attention mechanism to fine-tune the ability of the model to discern and prioritize critical features across the various omics datasets.

The novel confidence criterion also illustrates increased prediction performance compared to the conventional MCP strategy. The results consistently show that integrating the TCP criterion contributes to performance enhancements in most experiments. Specifically, on the BRCA dataset, the application of TCP results in significant improvements in ACC, AUC, and F1 scores (t-test $P < 0.05$). The ROSMAP and LGG datasets, including the confidence networks, also achieve significantly higher accuracy and F1 scores.

We further monitored the progression of training and testing losses over increasing epochs to investigate if the TCP-based confidence network can help reduce the risk of overfitting. Figure 3 illustrates the learning curve comparison of the ROSMAP classification task. It shows that the model without using ConfNet exhibits tendencies toward overfitting (around epoch 300), while a limitation is effectively reduced by including the TCP criterion. This suggests that the novel confidence criterion can improve prediction performance and contribute to the generalization capabilities of the model, making it more robust and reliable when applied to unseen data.

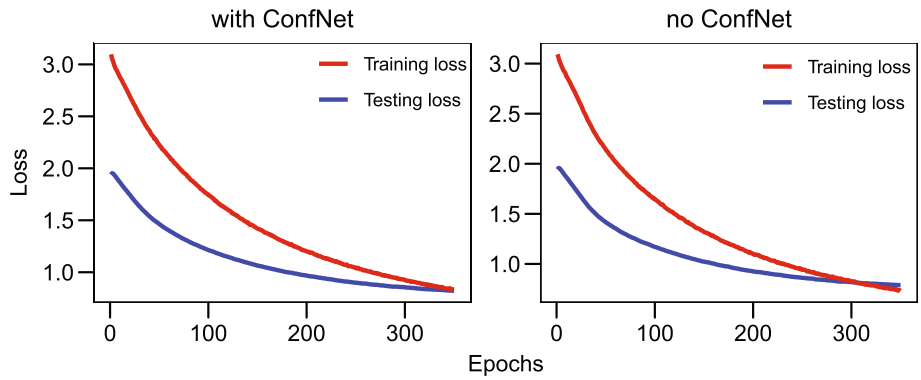

Fig. 3 Comparison of training and testing curves with and without the ConfNet in the ROSMAP classification task

Table 5 Top important biomarkers identified through our algorithm

Dataset	Omics	Top identified biomarkers
ROSMAP	mRNA	*NPNT, ANKRD30B, TCEA3, PRTN3, ZNF652-AS1, SAMD4A, SYTL1, AC131056.3, NRIP2*
	methy	*C10orf99, RORC, CRMP1, TMEM59, SNRPA, NGEF, C1orf83, TMEM85, ATP6V1B1, KIAA1267, HYAL2*
	miRNA	*hsa-miR-375, hsa-miR-767-5p, hsa-miR-146b-5p, hsa-miR-651, hsa-miR-93, hsa-miR-1266, hsa-let-7i, hsa-miR-224, hsa-miR-129-5p, hsa-miR-132, hsa-miR-330-3p*
BRCA	mRNA	*ZIC4, SCN7A, HPDL, PPP1R14C, SFRP1, TRIM29, WDR67, DUSP7, FABP7, PI3, CAMKV , CCDC150, CDKN2A, COG2, FANCE, NR2E1, PHOSPHO2, C1orf112, C9orf100, CCDC99, FANCB, GSG2, LBR, NUBPL, SGOL2, YBX1*
	methy	*COQ3, SOX21*
	miRNA	*hsa-mir-9-3, hsa-mir-374a, hsa-mir-92b*
LGG	mRNA	*LOC349196, TMEM179, BBC3, WDR53, ZNF77, ZSCAN16, GSTM3, LOC442308, LBX2, BPHL, APOL4, BTN2A2, DUSP10, GP9, HGF, IRGM, LOC222699, LYVE1, MSX2P1, TARSL2*
	methy	*SIGLEC11, GDF3, TWSG1, OR6Q1*
	miRNA	*hsa-mir-1234, hsa-mir-142, hsa-mir-21, hsa-mir-3655, hsa-mir-618, hsa-mir-9-3*

Identification of important biomarkers

The results of our biomarker identification experiments, focusing on mRNA expression, DNA methylation, and miRNA expression, are comprehensively detailed in Table 5. This table highlights the top thirty biomarkers identified from each dataset. To validate the relevance of these biomarkers, we cross-referenced them with existing medical literature. The inter-omics relevance among the top findings are depicted in Fig. 4. The top biomarkers exhibit significantly greater interactive effects than random tri-omics combinations, with t-test $P < 2.2\text{E-}16$ for the ROSMAP dataset and $P < 0.011$ for the BRCA dataset. These findings support that the proposed model effectively facilitates the identification of highly relevant biomarkers.

Several key findings emerged in the analysis of biomarkers within the BRCA dataset. The loss of *SFRP1* has been linked with the progression of breast cancer and a poorer prognosis in early-stage tumors [35]. Furthermore, *TRIM29* plays a role in suppressing *TWIST1* and the invasive behavior of breast cancer [36]. The expression of *C1ORF112* is notably high in both breast and cervical cancers [37]. Chen et al. [38] observed that the genetic depletion of *GSG2* marginally inhibits the growth of breast cancer cells while significantly enhancing their sensitivity to MLN8237 treatment. Additionally,

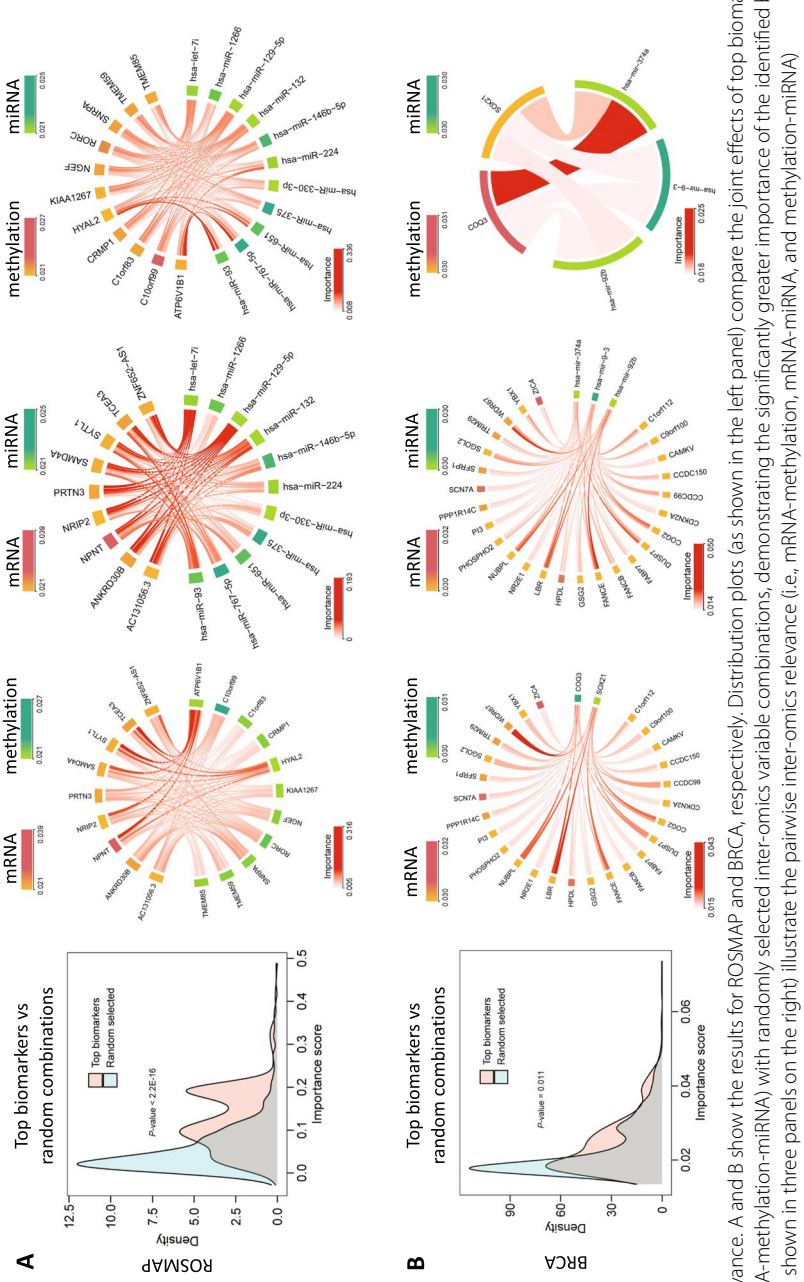

Fig. 4 Inter-omics relevance. A and B show the results for ROSMAP and BRCA, respectively. Distribution plots (as shown in the left panel) compare the joint effects of top biomarkers (three-omics combinations, i.e., mRNA-methylation-miRNA) with randomly selected inter-omics variable combinations, demonstrating the significantly greater importance of the identified biomarkers (t-test). The Chord diagrams (as shown in three panels on the right) illustrate the pairwise inter-omics relevance (i.e., mRNA-methylation, mRNA-miRNA, and methylation-miRNA)

the distribution of *miR-374a* in breast tumors has been examined by Li et al. [39], with implications for its role in breast cancer progression.

In the ROSMAP dataset, several biomarkers with significant implications for AD are discovered. For example, *NPNT* has been recognized as a crucial gene differentially expressed in brain tissues associated with late-onset AD [40]. Moreover, *PRTN3* has been identified as a key protective factor across various cognitive states, including dementia, mild cognitive impairment, and no cognitive impairment, and is instrumental in cognitive decline [40]. The role of *TMEM59* haploinsufficiency in reducing pathology and cognitive impairment has been well documented in the 5xFAD mouse model of AD [41]. Additionally, *Hsa-miR-375* has emerged as a novel circulating biomarker associated with extracellular vesicles in AD [42]. Furthermore, *Hsa-miR-132*, noted for its prosurvival, anti-inflammatory, and memory-enhancing functions in the nervous system, has been consistently observed to be downregulated in AD [43].

In analyzing biomarkers from the LGG dataset, several notable findings related to glioma cells have been observed. Li et al. [44] reported an increase in the expression of *GSTM3* in glioma cells compared to normal cells. Chen et al. [45] revealed that *LBX2-AS1*, a long non-coding RNA (lncRNA), is significantly upregulated in glioma, with its expression being associated with the prognosis of glioma patients. The role of *SIGLEC11* in maintaining microglia in a silent homeostatic status through sensing the intact glycocalyx of neighboring cells [46]. Zhang et al. (2020) [47] found that increased expression of *Sema3C*, which is regulated by *miR-142-5p*, indicates a poor prognosis in glioma. Additionally, Hermansen et al. (2013) [48] noted that *MiR-21* expression in the tumor cell compartment is associated with an unfavorable prognosis in gliomas.

These findings from our experiments align with existing research, thereby substantiating the robustness of our methodology in pinpointing biologically pertinent biomarkers critical for assessing disease impact.

Discussion

Our model is based on the existing shortcomings of multi-omics research, integrating auxiliary classifier-enhanced autoencoder, attention module, and the confidence network and verifying the rationality of these key components through argumentation and experimental comparison. The model not only demonstrated its state-of-the-art disease prediction ability on Alzheimer's disease, breast cancer, gliomas, and renal cancer but also successfully detected important biomarkers for understanding disease mechanisms through feature ablation experiments.

Our contribution mainly lies in three parts. Firstly, we designed auxiliary classifiers for each omics-specific autoencoder before combining omics data. These auxiliary classifiers help train autoencoders to accurately optimize sub-representations based on task requirements, better utilize the unique features present in each omics source, and thus improve classification performance. Secondly, the attention mechanism is an effective data fusion processing method, where the model can focus more attention on omics features that significantly contribute to the classification results, further optimizing the prediction performance. Finally, the TCP criterion evaluates model confidence by comparing the predicted probabilities with real labels, thereby effectively calibrating the overconfident predictions often observed in the standard softmax output.

Our model successfully pinpoints meaningful biomarkers within each omics data type in the BRCA, ROSMAP, and LGG datasets, demonstrating robust associations with various diseases. The biomarkers identified align closely with existing medical literature findings, reinforcing the biological significance of our discoveries. The congruence of our results with established literature not only validates the efficacy of our methodology but also emphasizes the potential of these biomarkers in clinical diagnosis and their contribution to the progression of various diseases.

While our model demonstrates impressive performance, there is potential for further enhancement. First, significant opportunities remain to delve into various data fusion methodologies. For example, utilizing interaction features over basic concatenation might yield more insightful revelations regarding the interplay among features from different modalities. Another aspect worthy of exploration is the differential contributions of various model components to prediction performance. Understanding the reasons behind these varying sensitivities in different components, particularly across a range of complex diseases, presents a valuable direction for future research.

Conclusion

In this study, we have developed a multi-omics data integration framework that significantly enhances the prediction accuracy of complex diseases and demonstrates stable prediction performance across various datasets. By adeptly compressing high-dimensional data, extracting key biologically relevant features, and further leveraging omics-specific classifiers along with true class probability optimization, our framework has demonstrated superior disease classification performance compared to SOTA methods. Rigorous validation across datasets confirms the robustness and effectiveness of our model, which also serves as a potent tool for identifying critical biomarkers. These biomarkers offer profound insights into disease diagnosis and the underlying mechanisms, potentially guiding the development of targeted therapies. Thus, our work is a significant stride toward advancing precision medicine and sets the stage for subsequent research to enhance disease prediction and treatment.

Authors' contributions
Study design, Xiaohui Yao and Shan Cong; methodology, Xiaohan Jiang, Xiaohui Yao and Shan Cong; data preparation, Haoran Luo; experiments: Xiaohan Jiang and Xiaohui Yao; writing-original draft, Xiaohan Jiang, Xiaohui Yao and Shan Cong; writing-review and editing, Hong Liang, Xiufen Ye and Yanhui Wei.

Funding
This work is supported by the National Key Research and Development Program of China (2022YFB4703500), the National Natural Science Foundation of China (62102115, 62103116), Shandong Provincial Natural Science Foundation (2022HWYQ-093), the Natural Science Foundation of Heilongjiang Province (LH2022F016), and the Fundamental Research Funds for the Central Universities (3072022TS2614).

Availability of data and materials
No datasets were generated or analysed during the current study.

Code availability
The code is available at https://github.com/Yaolab-fantastic/MOCAT.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
All authors have read and agreed to the published version of the manuscript.

Competing interests
The authors declare no competing interests.

Received: 21 December 2023 Accepted: 29 February 2024
Published online: 05 March 2024

References

1. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinforma Biol Insights. 2020;14:1177932219899051.
2. Kreitmaier P, Katsoula G, Zeggini E. Insights from multi-omics integration in complex disease primary tissues. Trends Genet. 2023;39(1):46–58. https://www.sciencedirect.com/science/article/pii/S0168952522002256.
3. Yan H, Bi L, Wang Y, Zhang X, Hou Z, Wang Q, et al. Integrative analysis of multi-omics data reveals distinct impacts of DDB1-CUL4 associated factors in human lung adenocarcinomas. Sci Rep. 2017;7(1):333.
4. Argelaguet R. Statistical methods for the integrative analysis of single-cell multi-omics data. 2021. https://www.repository.cam.ac.uk/handle/1810/315822. Accessed 16 July 2023.
5. Colomé-Tatché M, Theis FJ. Statistical single cell multi-omics integration. Curr Opin Syst Biol. 2018;7:54–59. https://www.sciencedirect.com/science/article/pii/S2452310018300039.Accessed 25 Aug 2023.
6. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. Brief Bioinform. 2021;23(1):bbab454.
7. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep. 2016;6:26094.
8. An N, Ding H, Yang J, Au R, Ang TFA. Deep ensemble learning for Alzheimer's disease classification. J Biomed Inform. 2020;105:103411.
9. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3(Mar):1157–82.
10. Poirion OB, Chaudhary K, Garmire LX. Deep Learning data integration for better risk stratification models of bladder cancer. AMIA Joint Summits Transl Sci Proc. 2018;2017:197–206. https://europepmc.org/articles/PMC5961799. Accessed 20 July 2023.
11. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. Clin Cancer Res. 2018;24(6):1248–59.
12. Baek B, Lee H. Prediction of survival and recurrence in patients with pancreatic cancer by integrating multi-omics data. Sci Rep. 2020;10(1):18951.
13. Wang X, Hu Z, Yu T, Wang Y, Wang R, Wei Y, et al. Con-AAE: contrastive cycle adversarial autoencoders for single-cell multi-omics alignment and integration. Bioinformatics. 2023;39(4):btad162.
14. Lin X, Tian T, Wei Z, Hakonarson H. Clustering of single-cell multi-omics data with a multimodal deep learning method. Nat Commun. 2022;13(1):7705.
15. Tan J, Ung M, Cheng C, Greene CS. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. Pac Symp Biocomput. 2015;20:132–43.
16. Chen L, Cai C, Chen V, Lu X. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. BMC Bioinformatics. 2016;17(1):97–107.
17. Khalili M, Majd HA, Khodakarim S, Ahadi B, Hamidpour M, Majd HA. Prediction of the thromboembolic syndrome: an application of artificial neural networks in gene expression data analysis. J Paramedical Sci (JPS) Spring. 2016;7:15–22.
18. Chen Q, Song X, Yamada H, Shibasaki R. Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016. Arizona: Phoenix; 2016. pages 338–44.
19. Schölkopf B, Platt JC, Hoffman T. Adv Neural Inf Process Syst. 2007;19:753–60.
20. Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun. 2021;12(1):3445.
21. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. Biotechnol Adv. 2021;49:107739.
22. Corbière C, Thome N, Bar-Hen A, Cord M, Pérez P. Addressing failure prediction by learning model confidence. Adv Neural Inf Process Syst. 2019;32:2898–909.
23. Simon R. Class probability estimation for medical studies. Biom J. 2014;56(4):597–600.
24. Fix E, Hodges JL. Discriminatory analysis. Nonparametric discrimination: Consistency properties. Int Stat Rev/Rev Int Stat. 1989;57(3):238–47.
25. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
26. Xie G, Dong C, Kong Y, Zhong JF, Li M, Wang K. Group lasso regularized deep learning for cancer prognosis from multi-omics and clinical features. Genes. 2019;10(3):240.
27. Ho TK. Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. vol. 1. IEEE; 1995. p. 278–282.
28. Chen T. Guestrin C. Xgboost: A scalable tree boosting system; 2016. p. 785–94.
29. Schwenker F, Trentin E. Pattern classification and clustering: A review of partially supervised learning approaches. Pattern Recogn Lett. 2014;37:4–14.
30. Van De Wiel MA, Lien TG, Verlaat W, van Wieringen WN, Wilting SM. Better prediction by use of co-data: adaptive group-regularized ridge regression. Stat Med. 2016;35(3):368–81.
31. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics. 2019;35(17):3055–62.

32. Hong D, Gao L, Yokoya N, Yao J, Chanussot J, Du Q, et al. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. IEEE Trans Geosci Remote Sens. 2020;59(5):4340–54.
33. Arevalo J, Solorio T, Montes-y Gómez M, González FA. Gated Multimodal Units for Information Fusion. 2017. arXiv preprint arXiv:1702.01992.
34. Han Z, Yang F, Huang J, Zhang C, Yao J. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 18-24, 2022, New Orleans, LA, USA. 2022. p. 20707–17.
35. Klopocki E, Kristiansen G, Wild PJ, Klaman I, Castanos-Velez E, Singer G, et al. Loss of SFRP1 is associated with breast cancer progression and poor prognosis in early stage tumors. Int J Oncol. 2004;25(3):641–9.
36. Ai L, Kim WJ, Alpay M, Tang M, Pardo CE, Hatakeyama S, et al. TRIM29 suppresses TWIST1 and invasive breast cancer behavior. Cancer Res. 2014;74(17):4875–87.
37. Edogbanya J, Tejada-Martinez D, Jones NJ, Jaiswal A, Bell S, Cordeiro R, et al. Evolution, structure and emerging roles of C1ORF112 in DNA replication, DNA damage responses, and cancer. Cell Mol Life Sci. 2021;78:4365–76.
38. Chen A, Wen S, Liu F, Zhang Z, Liu M, Wu Y, et al. CRISPR/Cas9 screening identifies a kinetochore-microtubule dependent mechanism for Aurora-A inhibitor resistance in breast cancer. Cancer Commun. 2021;41(2):121–39.
39. Li JY, Zhang Y, Zhang WH, Jia S, Kang Y, Tian R. Effects of differential distribution of microvessel density, possibly regulated by miR-374a, on breast cancer prognosis. Asian Pac J Cancer Prev. 2013;14(3):1715–20.
40. McCorkindale AN, Patrick E, Duce JA, Guennewig B, Sutherland GT. The Key Factors Predicting Dementia in Individuals with Alzheimer's Disease-Type Pathology. Front Aging Neurosci. 2022;14:831967.
41. Meng J, Han L, Zheng N, Xu H, Liu Z, Zhang X, et al. TMEM59 haploinsufficiency ameliorates the pathology and cognitive impairment in the 5xFAD mouse model of alzheimer's disease. Front Cell Dev Biol. 2020;8:596030.
42. Vogrinc D, Goričar K, Kunej T, Dolžan V. Systematic search for novel circulating biomarkers associated with extracellular vesicles in Alzheimer's disease: Combining literature screening and database mining approaches. J Personal Med. 2021;11(10):946.
43. Salta E, Sierksma A, Eynden EV, Strooper BD. miR-132 loss de-represses ITPKB and aggravates amyloid and TAU pathology in Alzheimer's brain. EMBO Mol Med. 2016;9(8):1005–18.
44. Li G, Cai Y, Wang C, Huang M, Chen J. LncRNA GAS5 regulates the proliferation, migration, invasion and apoptosis of brain glioma cells through targeting GSTM3 expression. The effect of LncRNA GAS5 on glioma cells. J Neuro-Oncol. 2019;143:525–36.
45. Chen Q, Gao J, Zhao Y, Hou R. Retracted artile: long non-coding RNA LBX2-AS1 enhances glioma proliferation through downregulating microRNA-491-5p. Cancer Cell Int. 2020;20:1–11.
46. Linnartz-Gerlach B, Kopatz J, Neumann H. Siglec functions of microglia. Glycobiology. 2014;24(9):794–9.
47. Zhang H, Ma H, Zhang W, Duan D, Zhu G, Cao W, et al. Increased expression of Sema3C indicates a poor prognosis and is regulated by miR-142-5p in glioma. Biol Pharm Bull. 2020;43(4):639–48.
48. Hermansen SK, Dahlrot RH, Nielsen BS, Hansen S, Kristensen BW. MiR-21 expression in the tumor cell compartment holds unfavorable prognostic value in gliomas. J Neuro-Oncol. 2013;111:71–81.

Publisher's Note