Multi-Level Confidence Learning for Trustworthy Multimodal Classification

Xiao Zheng[1], Chang Tang[2,*], Zhiguo Wan[3], Chengyu Hu[2], Wei Zhang[4]

[1]School of Computer, National University of Defense Technology, Changsha 410073, China
[2]School of Computer Science, China University of Geosciences, Wuhan 430074, China
[3]Zhejiang Lab, Hangzhou 311121, China
[4]Shandong Computer Science Center (National Supercomputing Center in Jinan), Jinan 250000, China

Abstract

With rapid development of various data acquisition technologies, more and more multimodal data come into being. It is important to integrate different modalities which are with high-dimensional features for boosting final multimodal data classification task. However, existing multimodal classification methods mainly focus on exploiting the complementary information of different modalities, while ignoring the learning confidence during information fusion. In this paper, we propose a trustworthy multimodal classification network via a multi-level confidence learning, referred to as MLCLNet. Considering that a large number of feature dimensions could not contribute to final classification performance but disturb the discriminability of different samples, we propose a feature confidence learning mechanism to suppress some redundant features, as well as enhancing the expression of discriminative feature dimensions in each modality. In order to capture the inherent sample structure information implied in each modality, we design a graph convolutional network branch to learn the corresponding structure preserved feature representation and generate modal-specific initial classification labels. Since samples from different modalities should share consistent labels, a cross-modal label fusion module is deployed to capture the label correlations of different modalities. In addition, motivated the ideally orthogonality of final fused label matrix, we design a label confidence loss to supervise the network for learning more separable data representations. To the best of our knowledge, MLCLNet is the first work which integrates both feature and label-level confidence learning for multimodal classification. Extensive experiments on four multimodal medical datasets are conducted to validate superior performance of MLCLNet when compared to other state-of-the-art methods.

Introduction

Multimodal data becomes more and more ubiquitous in past decades due to rapid development of diverse data acquisition technologies (Lahat, Adali, and Jutten 2015; Tang et al. 2018; Muhammad et al. 2021; Duro-Castano et al. 2021; Tang et al. 2022). With different modalities, a certain object or scene can be described more comprehensively (e.g., depth data obtained from depth video camera can capture complementary information for traditional RGB camera when a scene is with poor light condition (Liu, Zhang, and Wu 2022)). By exploring complementary information of different modalities, multimodal learning can often improve the final performance of certain tasks, which has been widely studied in various fields such as medical-diagnosis (Wang et al. 2014; Siejka-Zielińska et al. 2021; Zhou et al. 2021), classification (Gómez-Chova et al. 2015), video processing (Zhang and Wu 2022; Zhu et al. 2022) and information retrieval (Wei et al. 2020; Jing et al. 2022). Despite significant progress, most of existing multimodal learning models heavily rely on fusion strategies, which results in challenges of deployment for safety-critical applications such as computer-aided diagnosis. Therefore, trustworthy multimodal learning is important for practical applications.

As to multimodal learning, previous methods can be categorized into three classes: feature-level fusion, decision-level fusion and multi-level fusion. The former tries to jointly learn a unified representation from features of multiple modalities by certain mutual losses, the latter fuses the decision obtained from each modality and the middle integrates both modality-wise features and decisions. Either way, they all aim to explore the correlated and complementary information between different modalities for boosting final performance. In past decades, a large number of multimodal learning models have been put forward (McFee, Lanckriet, and Jebara 2011; Wang et al. 2012, 2015; Baltrušaitis, Ahuja, and Morency 2018; Wang 2021; Huang et al. 2021). Early proposed methods focus on aggregating the energy or information of multimodal spaces to reach better performance than a single modal (Wang et al. 2016). However, it is a challenge to capture the nonlinear distribution and complex structure of high-dimensional multimodal data. Due to the powerful capability of adaptively feature learning, deep neural network has also been widely used in the field of multimodal learning for feature learning and fusion (Choi and Lee 2019; Wu et al. 2020; Hong et al. 2020; Wang et al. 2021). By designing rational neural network architectures, joint representations of multiple modalities can be learned through appropriate objective functions that related to certain tasks.

Although great success have been obtained, stability and explainablity are not well guaranteed in most of existing multimodal learning methods since the informativeness of feature, modality, and decision are not jointly perceived,

*Corresponding author.

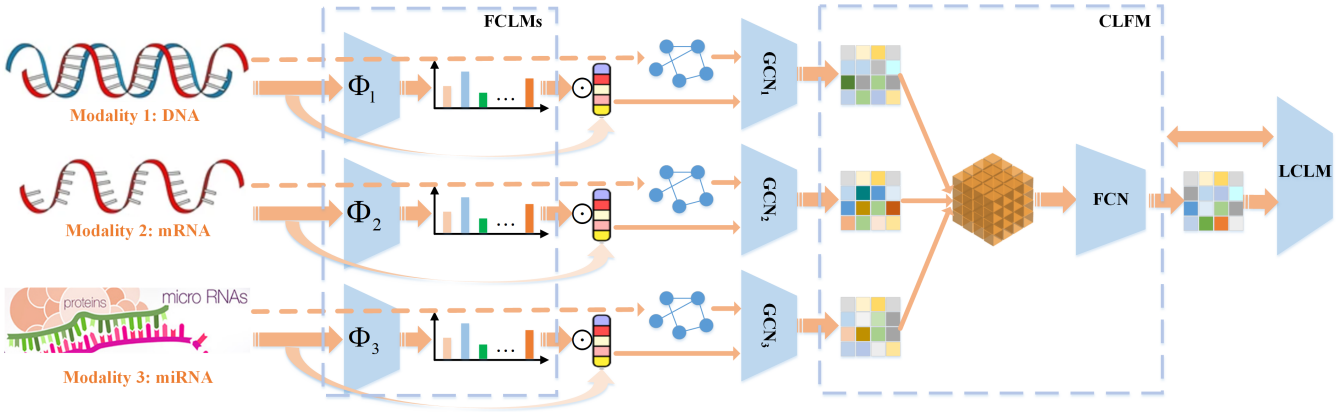

Figure 1: Framework of our proposed multi-level confidence learning network for trustworthy multimodal classification (we use 3 modal multi-omics data for example, and it is easy to extend the network to multiple modal case). The FCLM learns more discriminative feature representations for each modality by filtering out some redundant feature dimensions. The CLFM fuses initial predicted labels obtained from different modalities for capturing correlations in a cross-modal manner. Finally, the LCLM constructs a label confidence loss to supervised the network for generating more confident decisions. During feature representation learning, GCN is deployed for each modality to preserve the local geometrical structure of original data.

which results in unreliable results of the decision models. In multimodal data, uninformative features and trustless modalities often exist due to the unsatisfactory data collection process or instability of equipment (e.g., noisy RGB video modality under bad light condition (Garcia, Morerio, and Murino 2018; Lan et al. 2019), inherent noise of multiomics data (Yan et al. 2018; Argelaguet 2020) and sample missing phenomenon in certain modality/modalities (Ma et al. 2021b; Abdelaziz, Wang, and Elazab 2021)). Therefore, integrating the informativeness of each feature and each modality of different samples into multimodal learning is necessary and important for practical applications.

In this work, we propose a multi-level confidence learning network for trustworthy multimodal classification (MLCLNet), in which the informativeness of each feature in different modalities, modal certainty and label confidence are combined to enhance the trustworthiness of final decision. Specifically, we design a feature confidence learning mechanism to filter out the influence of some redundant features that disturb the discriminability of different samples. In order to fully exploit the inherent sample structure information implied in each modality, a graph convolutional network (GCN) branch is deployed for each modality to learn the corresponding structure preserved feature representation and generate modal-specific initial classification labels. Afterwards, a cross-modal label fusion module is developed to capture the label correlations and consistency of different modalities. In addition, considering the orthogonality of final fused idea label matrix, we invent a label confidence loss to supervise the network for learning more discriminative data representations. Figure 1 briefly shows the structure of our proposed MLCLNet. In a nutshell, the contributions of this work can be summarized as follows:

- To the best of our knowledge, we are the first to integrate both feature and label-level confidence learning for trustworthy multimodal classification;
- Different from previous methods which fuse information in low-level feature space, we design a label fusion module to exploit higher-level correlations of different modalities in the label space;
- With the fused label matrix, a label confidence loss is designed to supervise the network for learning more discriminative data representations to boost final classification performance. Extensive experiments on four multimodal medical datasets are conducted to validate the efficacy of the proposed MLCLNet.

Related Work

Multimodal learning has been widely investigated in past years and usually obtains remarkable performance improvement when compared to single modal learning models (Ramachandram and Taylor 2017; Baltrušaitis, Ahuja, and Morency 2018). The critical issue of multimodal learning is how to fuse complementary information from different modalities. Feature-level fusion methods directly integrate original data from multiple modalities, such as data concatenation or summation (Dalla Mura et al. 2015). However, it is hard to handle heterogeneous or non-consistent scale data using simple early fusion strategy. Decision-level fusion methods first generate prediction labels from each modality and then fuse the multiple outputs to get final decision (Wang et al. 2021). Multi-level fusion methods integrate both features and decisions from different modalities and therefore they can capture both low-level and high-level information (Poria, Cambria, and Gelbukh 2015; Hu and Singh 2021; Lee and van der Schaar 2021).

By taking the uncertainty of data and modality into consideration, trustworthy multimodal learning becomes a hot and cutting-edge topic. In order to capture both feature and modality informativeness, Han et al. (Han et al. 2022a) proposed a dynamical fusion network for trustworthy multimodal classification, in which a sparse gating is introduced

to capture the information variation of each within-modality feature and the true class probability is employed to assess the classification confidence of each modality. Wang et al. presented a novel multi-omics integrative method named MOGONET for biomedical classification (Wang et al. 2021), which jointly explores omics-specific learning and cross-omics correlation learning via view correlation discovery network (VCDN) (Wang et al. 2019) for effective multi-omics data classification. Since different types of omics data can provide unique class-level distinctiveness, rather than fusing features from different modalities, MOGONET uses VCDN to exploit the higher-level intra-view and cross-view correlations in the label space and get final trustworthy classification labels. In order to enhance confidence of prediction for diverse situations, Ma et al. (Ma et al. 2021a) introduced a novel Mixture of Normal-Inverse Gamma distributions (MoNIG) algorithm, in which the uncertainty of modalities can be efficiently estimated for adaptively modality integration to produce a trustworthy regression result. Based on Dempster-Shafer theory, Han et al. (Han et al. 2021, 2022b) proposed the variational Dirichlet for class probability distribution characterization, and different modalities were integrated at evidence level to consolidate the learning framework with both reliability and robustness against possible noise or corruption. In this paper, we propose a trustworthy multimodal classification network via both feature-level and label-level confidence learning.

Proposed MLCLNet

In this part, we give the detailed illustration of our proposed MLCLNet. Supposing we have N data samples described by M modalities and $\mathbf{x}_n^m$ denotes the n-th sample of the m-th modality, and $\mathbf{y}_n$ denotes the class label of the n-th sample. Multimodal classification aims to classify the N samples into C different classes by using the data from M modalities. Different from previous methods that fuse features or prediction labels, we design to adaptively learn feature confidence for filtering out some noisy/redundant feature dimensions for each sample in each modality, then a graph convolutional network is used to learn structure preserved feature representation and generate modality-specific class labels. Finally, a cross-modal label fusion module is deployed to capture the correlations of different modalities in high-level label space. In addition, the ideally orthogonality of final label matrix motivates us to design a label confidence loss for supervising the network to learn more discriminative data representations. Therefore, there are three main modules in MLCLNet, i.e., feature confidence learning module (FCLM), cross-modal label fusion module (CLFM) and label confidence learning module (LCLM), as shown by Figure 1. Following we will give detailed introduction of each module.

Feature Confidence Learning Module (FCLM)

It is well known that there is a large proportion of noisy/redundant features in high-dimensional data, which could not contribute to final learning performance but degenerate the discriminability of different samples. Therefore, how to reduce the influence of noisy/redundant features should be important for performance gain. In previous works, sparse regularization is a popular strategy for handling high-dimensional data. However, the informativeness of one feature for different samples are regarded as the same in traditional models, which is in conflict with actual situation. To this end, we design a FCLM for each modality to retain important features as well as suppress redundant features, which promotes the feature representation and trustworthiness within each modality.

Specifically, given a certain sample in the m-th modality $\mathbf{x}_m \in \mathbb{R}^{d_m}$ where d_m denotes the feature dimension of the m-th modality, we train an encoder network to learn its corresponding feature informativeness vector, i.e., $\mathbf{w}_m \in \mathbb{R}^{d_m}$. For simplicity, we use the sigmoid activation to scale the learned feature informativeness values, which can be mathematically formulated as follows:

$$\mathbf{w}_m = \sigma(\Phi(\mathbf{x}_m)), \tag{1}$$

where Φ represents the operations of encoder network and $\sigma(\cdot)$ refers to the sigmoid activation function. For high-dimensional data, it is useful to impose sparsity on the features to seek a small subset of relevant features. Therefore, l_0-norm regularization can be employed on w_m. However, it is hard to optimize l_0-norm regularization in practice, we use l_1-norm instead for easier solution. With the learned feature informativeness measure, the filtered features are obtained by the element-wise production between original features and $\mathbf{w}_m$ for each sample in each modality as follows:

$$\hat{\mathbf{x}}^m = \mathbf{x}^m \odot \mathbf{w}^m. \tag{2}$$

Cross-modal Label Fusion Module (CLFM)

For feature representation learning, the implied geometric structure of original data is a useful prior. With selected subset of relevant features, we use GCN for each modality to learn structure preserved compact feature representation.

For modality m, each sample is regarded as a node in the sample similarity graph. GCN aims to learn node features on the graph for classification tasks by aggregating both the features of each node and its neighbours characterized by the graph. For each GCN branch, there are two kinds of inputs, i.e., original feature matrix $\mathbf{X}_m \in \mathbb{R}^{N \times d_m}$ and the corresponding graph matrix $\mathbf{A}_m \in \mathbb{R}^{N \times N}$. In this work, we build each GCN by stacking a series of convolutional layers. In detail, each layer can be defined as:

$$\mathbf{H}_m^{l+1} = f(\mathbf{H}_m^l, \mathbf{A}_m) = \sigma(\mathbf{A}_m \mathbf{H}_m^l \mathbf{W}_m^l), \tag{3}$$

where $\mathbf{H}_m^l$ is the input of the l-th layer and $\mathbf{W}_m^l$ denotes the weight matrix of the l-th layer which needs to be learned during network training, and $\sigma(\cdot)$ also denotes the non-linear activation function.

Similar to previous work, the adjacency matrix $\mathbf{A}_m$ is also constructed by calculating the cosine similarity between pairs of nodes and we retain edges with cosine similarity larger than a threshold τ. Specifically, the adjacency between node i and node j in the graph can be calculated as follows:

$$\mathbf{A}_m(i,j) = \begin{cases} s(\mathbf{x}_m^i, \mathbf{x}_m^j), \; if \; i \neq j \; and \; s(\mathbf{x}_m^i, \mathbf{x}_m^j) \geq \tau \\ 0, \; otherwise \end{cases} \tag{4}$$

where $\mathbf{x}_m^i$ and $\mathbf{x}_m^j$ are the feature vectors of node i and node j, respectively. $s(\mathbf{x}_m^i, \mathbf{x}_m^j)$ is the cosine similarity between node i and j.

It should be noted that GCNs have been widely utilized in unsupervised and semi-supervised learning but seldom used for supervised classification tasks. In this work, based on previous learned feature representation, we extend GCNs to supervised classification and generate initial class labels for each modality. Given a set of training data $\mathbf{X}_{tr}$ which consists of N_{tr} data samples, we train a GCN with $\mathbf{X}_{tr}$ and the corresponding adjacency matrix $\mathbf{A}_{tr}$, and generate the classification prediction matrix $\mathbf{Y}_{tr} \in \mathbb{R}^{N_{tr} \times C}$ in which the i-th row represents the predicted label probability of the i-th training sample belonging to each class. For a new test sample $\mathbf{x}_{tr} \in \mathbb{R}^{d_m}$, the data matrix is extended as $\mathbf{X}_{trte} = [\mathbf{X}_{tr}; \mathbf{X}_{te}] \in \mathbb{R}^{(N_{tr}+1) \times d_m}$ and the corresponding adjacency matrix is extended as $\mathbf{A}_{trte} \in \mathbb{R}^{(N_{tr}+1) \times (N_{tr}+1)}$. Therefore, we can get final extended prediction label probability matrix as $\mathbf{Y}_{trte} \in \mathbb{R}^{(N_{tr}+1) \times C}$ in which the last row denotes the predicted label probability of the testing sample belonging to each class. In this way, the features of the test sample and the correlations between the test sample and training samples are both utilized for label prediction of the new test sample.

There are many previous multimodal learning models which fuse features from different modalities to generate a unified representation of original data. However, feature-level fusion is challenging and has no reliable guarantee when different modalities are heterogeneous. In addition, it is not easy to align different modalities in low-level feature space. In this work, instead of feature fusion, we try to exploit the high-level cross-modal correlations in the label space (Wang et al. 2021). For the m-th modality, the predicted label probability vector of the i-th sample is denoted as $\mathbf{y}_m^i \in \mathbb{R}^{1 \times C}$. We stack the predicted label probability vectors from different modalities to form a cross-modal label tensor as $\mathcal{T}^i \in \mathbb{R}^{C \times C \times \cdots \times C}$, and each entry of $\mathcal{T}^i$ can be calculated as:

$$\mathcal{T}^i(k_1, k_2, \cdots k_M) = y_{k_1}^{i,1} y_{k_2}^{i,2} \cdots y_{k_M}^{i,M}, \tag{5}$$

where $y_{k_m}^{i,m}$ denotes the mth entry of $\mathbf{y}_{k_m}^i$. Then, $\mathcal{T}^i$ is reshaped to a C^M dimensional vector and input to a fully connected layer with the output dimension of C. In this manner, the latent cross-modal label correlations can be well revealed to help improve the learning process and initial predictions from different modalities are integrated to generate final reliable decisions.

Label Confidence Learning Module (LCLM)

Although feature-level fusion and label-level fusion can effectively improve multimodal learning tasks, the final label confidence is often ignored and not well exploited for network supervision. In this work, we design a LCLM to measure the predicted label confidence and use it to supervise the network training. First, we recall the predicted label matrix $\mathbf{Y} \in \mathbb{R}^{N \times C}$ (for simplicity, we ignore the superscripts and subscripts to illustrate the concept), the p-th column of $\mathbf{Y}$ is $\mathbf{y}_p \in \mathbb{R}^{N \times 1}$, which consists of the probability of all data samples belonging to the p-th class.

Ideally, each sample should belong to only one class, i.e. each $\mathbf{y}_p$ is a one-hot vector (Huang, Gong, and Zhu 2020). However, this ideal condition is hard to reach in practical. In order to enable our proposed network towards this ideal case, we design a label confidence learning module by introducing a label guided objective loss, which is motivated by the natural property of label matrix. It is well known that we often constrain that the label matrix $\mathbf{Y}$ is orthogonal for clustering/classification tasks, i.e., $\mathbf{Y}^\top \mathbf{Y} = \mathbf{I}^{C \times C}$, which corresponds to the most confident prediction. Therefore, the question is *how to covert the orthogonality constraint to trainable loss function?*

Based on simple linear algebra, if two vectors are orthogonal, their inner product should be 0, which can be mathematically formulated as follows:

$$IP(\mathbf{y}_p, \mathbf{y}_q) = \mathbf{y}_p \cdot \mathbf{y}_q = 0, [p, q = 1, 2, \cdots, C]. \tag{6}$$

Therefore, if $IP(\mathbf{y}_p, \mathbf{y}_q)$ is with a large value, and the predicted p-th and q-th classes are not confident. For different columns of $\mathbf{Y}$, we can construct a label uncertainty matrix $\mathbf{U} \in \mathbb{R}^{C \times C}$ by using Eq. (6) on different set of all the cluster pairs. To this end, LCLM aims to minimise the matrix values (except the diagonal elements) of $\mathbf{U}$, which enforce the most confident classification results.

Note that Eq. (6) actually describes the correlation between two vectors buy using simple inner product operation, which is consistent with the widely used attention mechanism (Vaswani et al. 2017). Since $\mathbf{U}$ is a $C \times C$ matrix, we need to transform it to a scalar measure for network training. In traditional attention mechanism, attention is calculated on sample pairs. Therefore, we treat each class in $\mathbf{U}$ as a sample and reformulate original attention mechanism to suppress all the inter-class attention. For simplicity, we apply a softmax operation as self-attention to each class p and obtain a class pair correlation measurement as follows:

$$u(p, p') = \frac{\exp(\mathbf{U}(p, p'))}{\sum_{q=1}^{C} \exp(\mathbf{U}(p, q))}, \; p' \in [1, \cdots, C]. \tag{7}$$

Given Eq. (7), minimizing the class pair correlation is simplified into maximising $\{u(p, p)\}_{p=1}^{C}$.

Loss Functions

Since there are three major modules in our MLCLNet, i.e., FCLM, CLFM and LCLM, the final loss function also consists of three parts. Following we introduce the loss function for each module.

Feature Confidence Learning Loss As defined by Eq. (1), $\mathbf{w}_m$ is used to measure the informativeness of each feature dimension of each sample. Ideally, l_0-norm regularization can be used to select a subset of discriminative features but it is hard to solve. Therefore, we use l_1-norm regularization to $\mathbf{w}_m$ for sparsity approximation and lead to the feature confidence learning loss as follows:

$$\mathcal{L}_{FCL} = \sum\nolimits_{m=1}^{M} ||\mathbf{w}_m||_1. \tag{8}$$

Datasets		LGG			ROSMAP		
Method	Fusion stage	ACC	F1	AUC	ACC	F1	AUC
KNN	early	72.9±3.4	73.8±3.8	79.9±3.8	65.7±3.6	67.1±4.5	70.9±4.5
SVM	early	75.4±4.6	75.7±4.6	75.4±4.6	77.0±2.4	77.8±2.6	77.0±2.6
LR	early	76.1±1.8	76.7±2.7	82.3±2.7	69.4±3.7	73.0±3.5	77.0±3.5
RF	early	74.8±1.2	74.2±1.0	82.3±1.0	72.6±2.9	73.4±1.9	81.1±1.9
NN	early	73.7±2.3	74.8±3.7	81.0±3.7	75.5±2.1	76.4±2.5	82.7±2.5
GRridge	intermediate	74.6±3.8	75.6±4.4	82.6±4.4	76.0±3.4	76.9±2.3	84.1±2.3
BPLSDA	intermediate	75.9±2.5	73.8±2.3	82.5±2.3	74.2±2.4	75.5±2.5	83.0±2.5
BSPLSDA	intermediate	68.5±2.7	66.2±2.6	73.0±2.6	75.3±3.3	76.4±2.1	83.8±2.1
CF	intermediate	81.1±1.2	82.2±0.4	88.1±0.4	78.4±1.1	78.8±0.5	88.0±0.5
GMU	intermediate	80.3±1.5	80.8±1.2	88.6±1.2	77.6±2.5	78.4±1.6	86.9±1.6
MOGONET	decision	81.6±1.6	81.4±2.7	84.0±2.7	81.5±2.3	82.1±1.2	87.4±1.2
TMC	decision	81.9±0.8	81.5±0.4	87.1±0.4	82.5±0.9	82.3±0.6	88.5±0.6
MLCLNet	decision	**83.5±1.4**	**84.0±1.3**	**88.6±1.2**	**84.4±1.5**	**85.2±1.5**	**89.3±1.1**

Table 1: Classification results of different methods on the LGG and ROSMAP datasets (the best results are marked in bold font).

Cross-modal Label Fusion Loss Different to previous methods that fuse information of multiple modalities in the feature level, we fuse the cross-modal complementary information in the high-level label space. As shown in Eq. (5), a C^M dimensional label probability vector can be obtained by reshaping $\mathcal{T}^i$ for the i-th sample, then a fully connected layer with the output dimension of C can be used to generate final unified classification label $\hat{\mathbf{y}}^i$. For this module, the commonly used cross-entropy loss function is used for training supervision, which is formulated as follows:

$$\mathcal{L}_{FCL} = -\sum_{i=1}^{N_{tr}} \sum_{c=1}^{C} \hat{\mathbf{y}}_c^{\,i} \log \mathbf{y}_c^i, \tag{9}$$

where $\hat{\mathbf{y}}_c^{\,i}$ and $\mathbf{y}_c^i$ are the c-th element of $\hat{\mathbf{y}}^i$ and $\mathbf{y}^i$, respectively.

Label Confidence Learning Loss As shown in Eq. (7), maximising $\{u(p,p)\}_{p=1}^{C}$ could obtain the most confident prediction. If we treat $u(p,p)$ as the model prediction probability on the ground-truth class of a training sample, cross-entropy loss can be also exploited for training supervision, which can be formulated as:

$$\mathcal{L}_{LCL} = -\frac{1}{C}\sum_{p=1}^{C} \log u(p,p). \tag{10}$$

Finally, the overall objective function of MLCLNet can be obtained by combining above three terms as follows:

$$\mathcal{L} = \mathcal{L}_{FCL} + \lambda_1 \mathcal{L}_{CLF} + \lambda_2 \mathcal{L}_{LCL}, \tag{11}$$

where λ_1 and λ_2 are two weight parameters to balance different losses.

Experiments

In the section, we compare the proposed MLCLNet with some other state-of-the-art classification methods on four real-world multimodal datasets. Extensive experimental results validate the superiority of our propose network when compared with other counterparts. In addition, ablation studies are also conducted to demonstrate the effectiveness of different modules.

Experimental Settings

Datasets. Four benchmark multimodal medical datasets are used in our experiments, and the details of each dataset are as follows:

BRCA is used for breast invasive carcinoma PAM50 subtype classification, which contains 875 samples from 5 different classes.

LGG is used for grade classification in glioma, which contains 510 samples from 2 classes.

ROSMAP is used for Alzheimer's Disease diagnosis, which contains 351 samples of 2 classes (A Bennett et al. 2012; De Jager et al. 2018).

KIPAN is used for kidney cancer type classification, which contains 658 samples from 3 classes.

There are three different modalities associated in above datasets, i.e., mRNA expression, DNA methylation, and miRNA expression. BCRA, LGG, and KIPAN can be obtained from The Cancer Genome Atlas program (TCGA) [1].

Compared methods. In order to validate the superiority of the proposed MLCLNet, 12 competitors including 5

[1] https://www.cancer.gov/about-nci/organization/ccg/research/structuralgenomics/tcga

Datasets		BRCA			KIPAN		
Method	Fusion stage	ACC	WeightedF1	MacroF1	ACC	WeightedF1	MacroF1
KNN	early	74.2±2.4	73.0±2.5	68.2±2.5	96.7±1.1	96.7±1.1	96.0±1.4
SVM	early	72.9±1.8	70.2±1.7	64.0±1.7	99.5±0.3	99.5±0.3	99.4±0.4
LR	early	73.2±1.2	69.8±2.6	64.2±2.6	97.4±0.2	97.4±0.2	97.2±0.4
RF	early	75.4±0.9	73.3±1.3	64.9±1.3	98.1±0.6	98.1±0.6	97.5±1.1
NN	early	75.4±2.8	74.0±4.7	66.8±4.7	99.1±0.5	99.1±0.5	99.1±0.5
GRridge	intermediate	74.5±1.6	72.6±2.5	65.6±2.5	99.4±0.4	99.4±0.4	99.3±0.4
BPLSDA	intermediate	64.2±0.9	53.4±1.7	36.9±1.7	93.3±1.3	93.3±1.3	91.9±2.1
BSPLSDA	intermediate	63.9±0.8	52.2±2.2	35.1±2.2	91.9±1.2	91.8±1.3	89.5±1.4
CF	intermediate	81.5±0.8	81.5±0.9	77.1±0.9	99.2±0.5	99.2±0.5	98.8±0.9
GMU	intermediate	80.0±3.9	79.8±5.8	74.6±5.8	97.7±1.6	97.6±1.7	95.8±3.2
MOGONET	decision	82.9±1.8	82.5±1.7	77.4±1.7	**99.9±0.2**	**99.9±0.2**	**99.9±0.2**
TMC	decision	84.2±0.5	84.4±0.9	80.6±0.9	99.7±0.3	99.7±0.3	99.4±0.5
MLCLNet	decision	**86.4±1.6**	**87.8±1.7**	**82.6±1.8**	**99.9±0.7**	**99.9±0.2**	**99.9±0.2**

Table 2: Classification results of different methods on the BRCA and KIPAN datasets (the best results are marked in bold font).

single-modal methods and 7 multimodal classification models are used for comparison. For single-modal classification methods, data from different modalities are simply concatenated as input and these methods are: K-Nearest Neighbors (**KNN**) (Fix and Hodges 1989), Support Vector Machine (**SVM**) (Cortes and Vapnik 1995), l_1-norm regularized Linear Regression (**LR**), Random Forest (**RF**) (Ho 1995), and fully connected neural networks (**NN**). In addition, 7 multimodal classification models are as follows:

GRidge: Group-regularized (logistic) ridge regression (Van De Wiel et al. 2016), which makes structural use of multimodal data through group-specific penalties.

BPLSDA: Block partial least squares discriminant analysis (Singh et al. 2019), which explores multimodal data in latent space through discriminant analysis.

BSPLSDA: Block sparse partial least squares discriminant analysis (Singh et al. 2019), which is based on BPLSDA by selecting the most relevant features with sparse constraints.

CF: Concatenation of final multimodal representations (Hong et al. 2020), which integrates multiple modalities by concatenating late stage multimodal representations.

GMU: Gated multimodal units for information fusion (Ovalle et al. 2017), which generates an intermediate cross-modal representation based on the combination of data from different modalities.

TMC: Trusted multi-view classification (Han et al. 2021), which conducts decision fusion based on the confidence of different modalities.

MOGONET: Multiomics graph convolutional networks (Wang et al. 2021), which constructs a GCN for each modality for data structure preservation and captures cross-modal correlation via correlation discovery model.

Evaluation metrics. Different metrics are used to evaluate the performance of the compared methods. Since there are both binary classes and multiple classes in the used datasets, we use accuracy (ACC), F1 score (F1), and area under the receiver operating characteristic curve (AUC) for binary classification tasks evaluation, and we use accuracy (ACC), average F1 score weighted by support (F1_weighted), and macro-averaged F1 score (F1_macro) for multiple classes classification tasks.

Since each dataset needs to be partitioned into training part and testing part, similar to previous work (Wang et al. 2021; Han et al. 2022a), we run experiments 20 times and report the average results and standard deviation for avoiding bias of data partition. The Adam optimizer with learning rate decay is used for network training. For each time of experiment on each dataset, we stop the training process at 1200 epochs and output the testing results.

Experimental Results

Firstly, we compare the proposed MLCLNet with other methods on binary classification tasks. The detailed classification results of different methods on LGG and ROSMAP datasets are reported in Table 1. As can be seen from the results, the proposed MLCLNet achieves the best performance when compared with the other methods in terms of different metrics. Specifically, MLCLNet achieves 1.6% and 1.8% improvements over the second best results in terms of ACC and F1 on LGG dataset, and achieves 1.9%, and 2.9% improvements over the second best results in terms of ACC and F1 on ROSMAP dataset. When compared to single-modal classification methods, our proposed method obtains significantly improvements in terms of all metrics on different datasets.

Secondly, we further compare the proposed MLCLNet with other methods on multi-classification tasks and present the classification results of different methods on BRCA and KIPAN datasets in Table 2. From the experimental results, we can get the following observations: 1) The proposed MLCLNet can get the unique best results in terms of all metrics on BRCA dataset; 2) The proposed method consistently obtains better results than single-modal classification methods on the two datasets; 3) When compared to MOGONET which also fuses modal-wise predictions, MLCLNet still has prominent advantage.

In order to give an intuitive presentation of performance varying with the training process, we plot the values of training loss and different metrics of MLCLNet on BRCA dataset with varying training iteration epochs in Figure 2. As can be seen, our proposed MLCLNet can converges well within 1000 iteration epochs and get stable classification results.

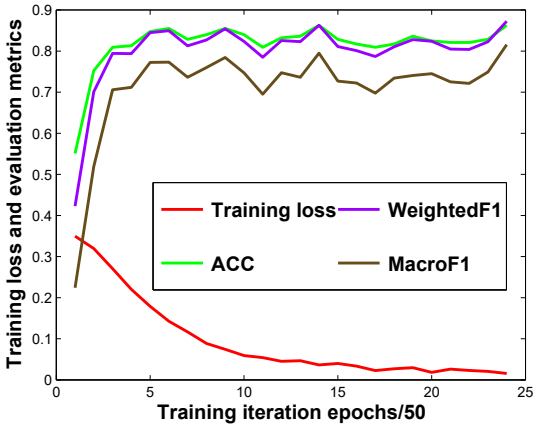

Figure 2: The training loss and the variation of different performance evaluation metrics on BRCA dataset.

Ablation Study

As elaborated in previous sections, there are three main modules in our proposed MLCLNet. In order to validate the efficacy of each module, we perform ablation study experiments with different settings of the network structure. Firstly, we remove the feature confidence learning module and keep other parts of MLCLNet, we call this variant MLCLNet_noFCL. Secondly, rather than constructing the cross-modal label correlation discovery tensor, we concatenate the label vector obtained from each modality and input to the fully connected layer for generating final C-dimensional prediction, and this version is called MLCLNet_noCLF. Finally, we remove the label confidence loss and regard the network as MLCLNet_noLCL. In Table 3, we show the final classification results of our proposed MLCLNet under different settings. The following observations can be obtained from the ablation experimental results: 1) All of the different modules contribute to the final performance gain; 2) The performance of MLCLNet degrades the most when the cross-modal label fusion module is removed, which verifies the importance of modality information fusion; 3) With the label confidence learning module, MLCLNet can also obtain significant performance improvement.

Dataset	Networks	ACC	F1
LGG	MLCLNet_noFCL	82.7±1.2	83.5±1.2
	MLCLNet_noCLF	80.3±1.3	81.2±1.3
	MLCLNet_noLCL	81.4±1.2	82.4±1.1
	MLCLNet	83.5±1.4	84.0±1.3
ROSMAP	MLCLNet_noFCL	83.7±1.4	84.3±1.3
	MLCLNet_noCLF	81.5±1.3	82.1±1.2
	MLCLNet_noLCL	82.0±1.2	83.6±1.4
	MLCLNet	84.4±1.5	85.2±1.5
BRCA	MLCLNet_noFCL	85.3±1.4	87.1±1.6
	MLCLNet_noCLF	82.7±1.5	83.4±1.6
	MLCLNet_noLCL	83.8±1.3	84.9±1.5
	MLCLNet	86.4±1.6	87.8±1.7
KIPAN	MLCLNet_noFCL	99.3±0.9	99.4±0.3
	MLCLNet_noCLF	98.3±0.8	98.9±0.2
	MLCLNet_noLCL	98.9±0.6	98.7±0.3
	MLCLNet	99.8±0.7	99.9±0.2

Table 3: Ablation study of the proposed MLCLNet on different datasets.

Conclusions

In this paper, we present a trustworthy multimodal classification network via multi-level confidence learning, named MLCLNet. Three main modules including feature confidence learning, cross-modal label fusion and label confidence learning are designed and integrated into MLCLNet for trustworthy feature representation learning and classification label prediction. Four practical medical multi-omics datasets are used to validate the efficacy of the proposed network and experimental results also demonstrate the superiority of MLCLNet when compared with other competitors. In addition, ablation studies are also conducted to verify the usefulness of different modules.

Acknowledgments

This work was supported in part by NSFC (No. 62076228 and 62073300), and in part by Key Research Project of Zhejiang Lab (No. K2022PD1BB01), and in part by the Natural Natural Science Foundation of Shandong Province (No. ZR2021LZH001). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for computation acceleration for this research.

References

A Bennett, D.; A Schneider, J.; Arvanitakis, Z.; and S Wilson, R. 2012. Overview and findings from the religious orders study. *Current Alzheimer Research*, 9(6): 628–645.

Abdelaziz, M.; Wang, T.; and Elazab, A. 2021. Alzheimers disease diagnosis framework from incomplete multimodal data using convolutional neural networks. *Journal of Biomedical Informatics*, 121: 103863.

Argelaguet, R. 2020. *Statistical methods for the integrative analysis of single-cell multi-omics data*. Ph.D. thesis, University of Cambridge.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.

Choi, J.-H.; and Lee, J.-S. 2019. EmbraceNet: A robust deep learning architecture for multimodal classification. *Information Fusion*, 51: 259–270.

Cortes, C.; and Vapnik, V. 1995. Support-vector networks. *Machine learning*, 20(3): 273–297.

Dalla Mura, M.; Prasad, S.; Pacifici, F.; Gamba, P.; Chanussot, J.; and Benediktsson, J. A. 2015. Challenges and opportunities of multimodality and data fusion in remote sensing. *Proceedings of the IEEE*, 103(9): 1585–1601.

De Jager, P. L.; Ma, Y.; McCabe, C.; Xu, J.; Vardarajan, B. N.; Felsky, D.; Klein, H.-U.; White, C. C.; Peters, M. A.; Lodgson, B.; et al. 2018. A multi-omic atlas of the human frontal cortex for aging and Alzheimers disease research. *Scientific data*, 5(1): 1–13.

Duro-Castano, A.; Borrás, C.; Herranz-Pérez, V.; Blanco-Gandía, M. C.; Conejos-Sánchez, I.; Armiñán, A.; Mas-Bargues, C.; Inglés, M.; Miñarro, J.; Rodríguez-Arias, M.; et al. 2021. Targeting Alzheimers disease with multimodal polypeptide-based nanoconjugates. *Science Advances*, 7(13): eabf9180.

Fix, E.; and Hodges, J. L. 1989. Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3): 238–247.

Garcia, N. C.; Morerio, P.; and Murino, V. 2018. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 103–118.

Gómez-Chova, L.; Tuia, D.; Moser, G.; and Camps-Valls, G. 2015. Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE*, 103(9): 1560–1584.

Han, Z.; Yang, F.; Huang, J.; Zhang, C.; and Yao, J. 2022a. Multimodal Dynamics: Dynamical Fusion for Trustworthy Multimodal Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20707–20717.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2021. Trusted multi-view classification. In *International Conference on Learning Representations*.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2022b. Trusted Multi-View Classification with Dynamic Evidential Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ho, T. K. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, 278–282. IEEE.

Hong, D.; Gao, L.; Yokoya, N.; Yao, J.; Chanussot, J.; Du, Q.; and Zhang, B. 2020. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5): 4340–4354.

Hu, R.; and Singh, A. 2021. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1439–1449.

Huang, J.; Gong, S.; and Zhu, X. 2020. Deep semantic clustering by partition confidence maximisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8849–8858.

Huang, Y.; Du, C.; Xue, Z.; Chen, X.; Zhao, H.; and Huang, L. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34: 10944–10956.

Jing, T.; Xia, H.; Hamm, J.; and Ding, Z. 2022. Augmented Multi-Modality Fusion for Generalized Zero-Shot Sketch-based Visual Retrieval. *IEEE Transactions on Image Processing*.

Lahat, D.; Adali, T.; and Jutten, C. 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9): 1449–1477.

Lan, X.; Ye, M.; Shao, R.; Zhong, B.; Yuen, P. C.; and Zhou, H. 2019. Learning modality-consistency feature templates: A robust RGB-infrared tracking system. *IEEE Transactions on Industrial Electronics*, 66(12): 9887–9897.

Lee, C.; and van der Schaar, M. 2021. A variational information bottleneck approach to multi-omics data integration. In *International Conference on Artificial Intelligence and Statistics*, 1513–1521. PMLR.

Liu, D.; Zhang, L.; and Wu, Y. 2022. LD-ConGR: A Large RGB-D Video Dataset for Long-Distance Continuous Gesture Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3304–3312.

Ma, H.; Han, Z.; Zhang, C.; Fu, H.; Zhou, J. T.; and Hu, Q. 2021a. Trustworthy Multimodal Regression with Mixture of Normal-inverse Gamma Distributions. *Advances in Neural Information Processing Systems*, 34: 6881–6893.

Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; and Peng, X. 2021b. SMIL: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2302–2310.

McFee, B.; Lanckriet, G.; and Jebara, T. 2011. Learning Multi-modal Similarity. *Journal of machine learning research*, 12(2).

Muhammad, G.; Alshehri, F.; Karray, F.; El Saddik, A.; Alsulaiman, M.; and Falk, T. H. 2021. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Information Fusion*, 76: 355–375.

Ovalle, J. E. A.; Solorio, T.; Montes-y Gómez, M.; and González, F. A. 2017. Gated Multimodal Units for Information Fusion. In *ICLR (Workshop)*.

Poria, S.; Cambria, E.; and Gelbukh, A. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2539–2544.

Ramachandram, D.; and Taylor, G. W. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6): 96–108.

Siejka-Zielińska, P.; Cheng, J.; Jackson, F.; Liu, Y.; Soonawalla, Z.; Reddy, S.; Silva, M.; Puta, L.; McCain, M. V.; Culver, E. L.; et al. 2021. Cell-free DNA TAPS provides multimodal information for early cancer detection. *Science advances*, 7(36): eabh0534.

Singh, A.; Shannon, C. P.; Gautier, B.; Rohart, F.; Vacher, M.; Tebbutt, S. J.; and Lê Cao, K.-A. 2019. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17): 3055–3062.

Tang, C.; Zheng, X.; Liu, X.; Zhang, W.; Zhang, J.; Xiong, J.; and Wang, L. 2022. Cross-view locality preserved diversity and consensus learning for multi-view unsupervised feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 34(10): 4705–4716.

Tang, C.; Zhu, X.; Liu, X.; Li, M.; Wang, P.; Zhang, C.; and Wang, L. 2018. Learning a joint affinity graph for multiview subspace clustering. *IEEE Transactions on Multimedia*, 21(7): 1724–1736.

Van De Wiel, M. A.; Lien, T. G.; Verlaat, W.; van Wieringen, W. N.; and Wilting, S. M. 2016. Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in medicine*, 35(3): 368–381.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, B.; Mezlini, A. M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; and Goldenberg, A. 2014. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3): 333–337.

Wang, L.; Ding, Z.; Tao, Z.; Liu, Y.; and Fu, Y. 2019. Generative multi-view human action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6212–6221.

Wang, M.; Li, H.; Tao, D.; Lu, K.; and Wu, X. 2012. Multimodal graph-based reranking for web image search. *IEEE transactions on image processing*, 21(11): 4649–4661.

Wang, T.; Shao, W.; Huang, Z.; Tang, H.; Zhang, J.; Ding, Z.; and Huang, K. 2021. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, 12(1): 1–13.

Wang, Y. 2021. Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1s): 1–25.

Wang, Y.; Lin, X.; Wu, L.; Zhang, W.; and Zhang, Q. 2015. Lbmch: Learning bridging mapping for cross-modal hashing. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 999–1002.

Wang, Y.; Wenjie, Z.; Wu, L.; Lin, X.; Fang, M.; and Pan, S. 2016. Iterative views agreement: an iterative low-rank based structured optimization method to multi-view spectral clustering. In *International Joint Conference on Artificial Intelligence*, 2153–2159. Association for the Advancement of Artificial Intelligence (AAAI).

Wei, X.; Zhang, T.; Li, Y.; Zhang, Y.; and Wu, F. 2020. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10941–10950.

Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; and Yang, Z. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*, 322–339. Springer.

Yan, J.; Risacher, S. L.; Shen, L.; and Saykin, A. J. 2018. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in bioinformatics*, 19(6): 1370–1381.

Zhang, X.; and Wu, X. 2022. Multi-modality deep restoration of extremely compressed face videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, J.; Zhang, X.; Zhu, Z.; Lan, X.; Fu, L.; Wang, H.; and Wen, H. 2021. Cohesive multi-modality feature learning and fusion for COVID-19 patient severity prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5): 2535–2549.

Zhu, X.; Zhu, Y.; Wang, H.; Wen, H.; Yan, Y.; and Liu, P. 2022. Skeleton Sequence and RGB Frame Based Multi-Modality Feature Fusion Network for Action Recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(3): 1–24.