Dual trustworthy mechanism for cross-domain multi-modality classification

*Note: Sub-titles are not captured in Xplore and should not be used

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—Multi-modality classification has become flourished in recent years. Traditional methods mainly focus on advancing deep neural networks (DNN) to meet high performance. However, the interpretability of these methods remains blind due to the complexity and ambiguity of DNN, which also causes distrust. This problem is enlarged in sensitive areas, such as biomedical computing and education evaluation. Hence, we propose a novel dual trustworthy mechanism for multi-modality classification (DTMC), which can make the process and results of DNN more credible and interpretable while increasing performance. Specifically, a confidence attention mechanism is performed from local and global views to improve the process' confidence by evaluating the attention scores and distinguishing the abnormal information. A confidence probability mechanism from local and global perspectives is conducted in the prediction stage to enhance the results' confidence. Additionally, this work concerted collect a multi-modality dataset for automated lesson plan grading. Extensive experiments on multi-modality medical classification datasets and lesson plan grading datasets show superior performance with the interpretability of the proposed method compared to the state-of-the-art (SOTA) methods. Our resources are open at https://github.com/ghh1125/data.**

***Index Terms*—Multi-modality learning, Trustworthiness and Interpretability learning, Cross-domain classification**

I. Introduction

The explosion of multi-modality data has increased the research interests in multi-modality classification [1], [2]. Multi-modality classification requires to mine related information from heterogeneous data to assign labels to samples. For instance, in biomedical computing, people expect to use different views of DNA features to infer diseases. While in education evaluation, multi-modality lesson plans containing pictures and text need to be scored reasonably by exploring them. The deep learning (DL) based methods like [3], [4], [5], [6], [7], [8] are wide use and show incredible performance. However, like a black box, the process of most DL methods remains unknown, making it difficult to persuade the public, especially when facing safety- and fairness-related tasks, as depicted in Fig. 1.

Due to the domain and component heterogeneity of different modalities, the fusion of modalities is tricky. Previous DL methods like CF [9] use simple concatenation for complex multi-modality information, and others use weighted summation, feature multiplication, or gating mechanism [10]. These methods need to be more comprehensive to fuse multi-modality information. Because for every single sample, there is an unbalanced informativeness between different modalities and features. HMCAN [11], [12], and [13] introduce a variety of variant attention mechanisms that are expected to obtain better multi-modality fusion capabilities. The attention mechanism can find adequate information from more fine-grained features, so attention can be a helpful tool, but it cannot always run well. Some distinctive features can attract more attention, but it does not mean the informativeness of these features, which we call $sharp - features$. Later work [14] dynamically fuse the feature and modality informativeness to promote accuracy, but the model confidence has yet to be further explored.

Considering the model's credibility in the real scene, [15] use Dirichlet distribution to model a distribution with evidence-level features to provide reliable uncertainty estimations. Dynamics [16] transfer the concept of True class probability from ConfidNet [17] to enhance the trust. However, it still needs further exploration to reach public acceptance. This work devotes to increasing the trustworthiness and interpretability of multi-modality classification. The proposed method designed two trustworthy mechanisms, the confidence

Identify applicable funding agency here. If none, delete this.

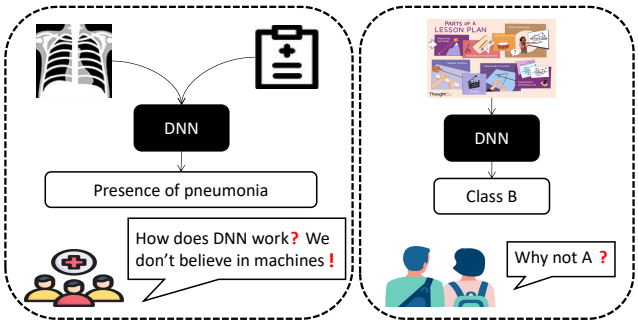

Fig. 1. The typical examples of distrust of deep neural networks.

attention module (CAM) and the conference probability module (CPM), to make the process and results of DNN more credible. Additionally, we notice that high attention scores of multi-modality features ($sharp - features$) do not represent high informativeness and alleviate this phenomenon by a penalty term.

Our contributions are summarized as follows:

- We propose a novel dual trustworthy mechanism called DTMC consisting of CAM and CPM. CAM performs sample-level multi-modality feature fusion and representation with trustworthy attention from a global and local perspective. CPM improves classification decision accuracy by reevaluating the predictive probabilities.
- This work deliberately collects a multi-modality dataset containing a range of lesson plans and opens it for community use. Moreover, we performed a series of operations for high-quality data, including data cleaning, feature extraction, et al.
- Wide experiments conducted on several popular datasets show DTMC performs result improvements compared with SOTA methods and provides a credible and interpretable process to make the prediction more solid.

The remains of the paper are structured as follows: Section 2 reviews related works briefly. Section 3 provides a detailed description of the proposed method. The experimental results are shown in Section 4. Finally, we conclude this paper and suggest future work in Section 5.

II. Related Work

Two main fields related to this work are multi-modality learning and trustworthy learning. This section briefly reviews the work on both fronts.

A. Multi-modality learning

Recently, multi-modality learning has become a research hotspot with the increase of multi-modality data such as advertisements and publications [18]. Generally speaking, multi-modality data refers to heterogeneous data such as text, pictures, and audio, but there is still no standard definition. For example, in disease classification, mRNA, DNA methylation, and miRNA expression data are considered three modalities. Tasks in multi-modality learning mainly comprise finding some unified representation of multi-modality information (representation), mapping a modality to another (translation), finding relationships between modality sub-components (alignment), integrating information (fusion), assisting scarce modalities by leveraging the knowledge of plentiful modalities (co-learning).

Previous work such as [10] and [19] et al. are excellent for small-scale multi-modality classification tasks, while [20] and [21] adapted to large-scale problems. The birth of the large-scale pre-training model [22] helps construct modalities data features upstream. In subsequent information processing, cross-modality feature fusion is crucial for modalities classification. Earlier work includes [19], [1] et al., using simple concatenate strategies. Other approaches use decision-making and dynamic fusion methods. With the popularity of the Attention mechanism [23], more and more methods tend to use the attention mechanism to integrate multi-modality data better.

B. Trustworthy learning

Research on trustworthy learning for DNN [24], [25], [26] [27] have always been prosperous. [24] give a complete theoretical treatment of the link between Gaussian processes and dropout and develop the tools necessary to represent uncertainty in deep learning. [28] pointed out that the confidence of most deep learning models has not been perfectly corrected, and the overall tendency is Over-Confident, which means the average confidence of the prediction is higher than the average accuracy of the prediction. [17] proposed the True Class Probability (TCP) concept, effectively enhancing the model's confidence.

Applications to trustworthy learning have also emerged in work such as [29] and [30]. [29] adopts normalized cross-entropy (NCE) loss to measure the quality of confidence score, while [30] uses a bi-directional approach for lattices (BiLatRNN) for confidence estimation. [31] noted that the common attention mechanism could not achieve credibility and proposed saliency-based explanations to achieve success.

III. Proposed Method

In this part, we first describe the datasets. Then introduce the proposed method in detail, including the CAM and CPM. The framework of the proposed method is shown in Fig. 2, and the detailed structure of each module can be referred to in Fig. 3.

A. Datasets

This study adopts several multi-modality datasets from biomedical computing and education evaluation, which are safety- and fairness-related for experiments in the credible and interpretable fusion and representation of heterogeneous data features.

a) Biomedical computing: four popular common datasets, ROSMAP, LGG, KIPAN, and BRCA, are tested. They all contain mRNA expression data, miRNA expression data, and DNA methylation data (meth). The ROSMAP dataset is used to classify Alzheimer's disease patients from

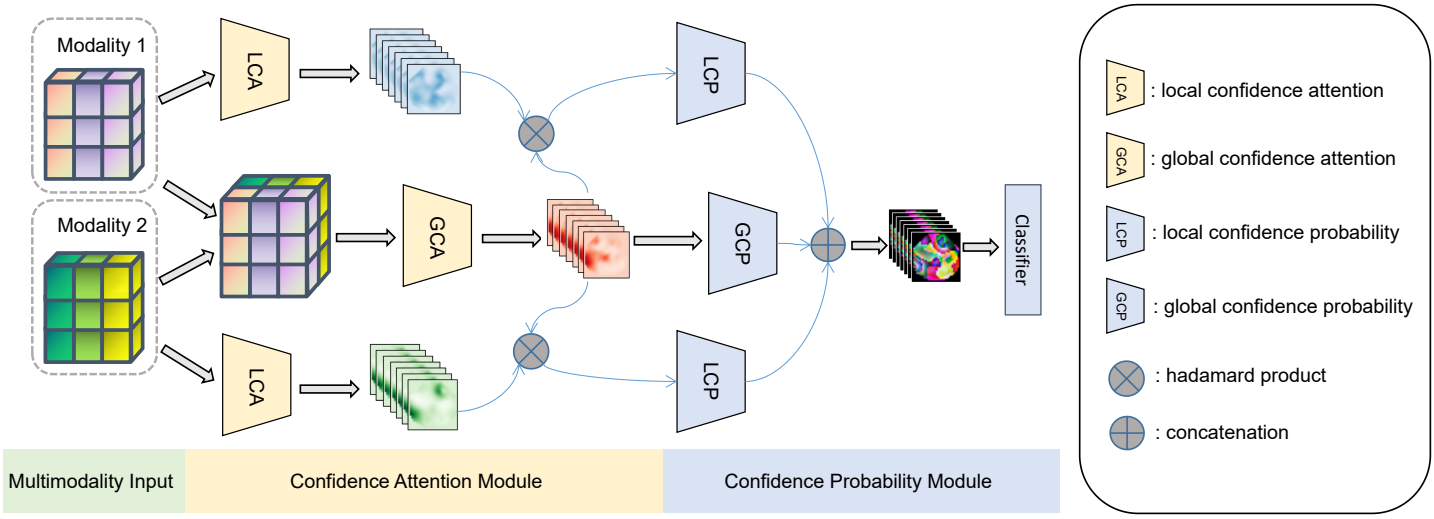

Fig. 2. The proposed method (two modalities example).

TABLE I
DETAILS OF THE SIX DATASETS. THE FOUR DATASETS IN THE UPPER PART ARE ABOUT BIOMEDICAL COMPUTING, AND THE LOWER TWO ARE ABOUT EDUCATION EVALUATION.

Datasets	Samples	Classes	Modalities
BRCA	875	5	3
KIPAN	658	3	3
LGG	510	2	3
ROSMAP	351	2	3
MATH	1615	3	2
ENGLISH	867	3	2

normal control patients, including 200 mRNA, 200 miRNA, and 200 meth features. The LGG dataset is for classifying low-grade gliomas (LGG), containing 2000 mRNA, 2000 miRNA, and 548 meth features. The KIPAN dataset is used for kidney cancer type classification with 2000 mRNA, 2000 miRNA, and 445 meth features. The BRCA dataset is about breast cancer (BRCA) PAM50 subtype classification, embracing 1000 mRNA, 1000 miRNA, and 503 meth features. The specific information can refer to the upper part of TABLE I.

b) Education evaluation: more than 70,000 lesson plans are manually collected. These lesson plans are designed by normal students and graded by internship and college teachers. The dataset contains multiple grades (elementary school and high school et al.) and subjects (chemistry, geography, and Physical Education et al.) lesson plans, so it needs to be classified by similar grades and subjects first. 17,773 samples remain after removing samples that are missing scores, missing files, and low-quality (incomplete modalities, the score gap given by the two reviews is large et al.).

Each sample has six items, including 'id,' 'text,' 'image,' 'structure,' 'label,' and 'subjects.' 'id' is the sample number after the privacy information is removed. 'Text' and 'image' are vectors from pre-trained models. Before pre-training, 512 words are extracted according to course keywords to unify the text since most texts consist of thousands of words in length. The specific method is to find ten common keywords with the highest frequency for samples under a subject, take a step size of 18 words forward and backward for every keyword, get them as input, and fill the parts less than 512 with the particular token 'blank.' Due to the different sizes of the images, we cut the pictures into a size of 100×100 pixels to form a patch sequence. Considering that some edge patches will overlap with internal patches, we use different symbols to distinguish them. The text and image sequence can be expressed as follows:

$$Txt_{\text{seq}} = \left\{ t_{key_1^1}, \cdots, t_{key_1^{18}}, t_{key_1^*}, \ldots, t_{key_1^{36}}, t_{key_2^1}, \cdots \right\}, \tag{1}$$

$$Img_{\text{seq}} = \{ p_{clc_1}, p_{clc_2}, \ldots, p_{sep}, p_{cov_1}, p_{cov_2}, \ldots \}. \tag{2}$$

where $t_{key_1^*}$ is a token of keyword, $t_{key_1^1}$ is the token of word obtained according to the step size. p_{clc_1} is a picture intercepted without coverage, p_{cov_1} is a picture intercepted with coverage, and p_{sep} is used as a mark to separate them.

We also collected a multi-hot vector called 'structure,' which contains eight necessary modules for lesson plans proposed by Chinese education experts, including 'textbook analysis,' 'learning situation analysis,' 'teaching objectives,' 'teaching priorities', 'teaching methods,' 'teaching tools,' 'teaching processes', and 'teaching reflections.' However, since it is not international-wide, we did not use it. The original 'label' of the lesson plan is $0-100$ points and is re-divided into three grades, including A, B, and C, corresponding to $100-90$, $89-80$, and $79-0$ points. 'subjects' is the subject corresponding to the sample. In this research, we select Math and English classes

for verification. The lower part of TABLE I shows the specific information.

B. Preliminaries

Suppose a dataset D with N samples in multi-modality classification is expressed as $D = \{X_n, Y_n\}_{n=1}^{N}$, for each sample X with M ($M > 1$) modalities can be expressed as $X = \{x_m\}_{m=1}^{M}$, where each x represents features of a modality, which generally be high-dimensional. The corresponding Y should be a binary or multivariate vector, depending on the number of classification labels. The multi-modality classification task aims to find a function f to map X and Y. Generally, f can be written as $f : X- > Y$.

Before the process of CAM, the raw data needs to be preprocessed by the feature extractor E and then get the w as the input of DTMC. For Biomedical computing tasks, E is a fully connected layer. While for Education evaluation tasks with complex features, E is the large-scale pre-trained model, which can often provide powerful assistance for downstream tasks [22] such as BERT and ResNet-50. In multi-modality classification, each modality x_m has a corresponding feature extractor, which can be expressed as the following formula:

$$w_m = \sigma(E_m(x_m)), \tag{3}$$

where σ is the activation function and E_m is the feature extractor.

C. Confidence Attention Module

Traditional attention mechanisms fail to realize that high attention does not always represent informativeness, which means the excessive participation of some invalid or harmful information, threatening the downstream network. It is, therefore, essential to improve the credibility of the attention mechanism. The CAM mainly includes Local Confidence Attention (LCA) and Global Confidence Attention (GCA) to learn multi-modality information from local and global perspectives with a confidence evaluation to enhance the reliability of the attention mechanism.

a) Local Confidence Attention: The input of every modality often contains some noise or useless features. The attention mechanism is expected to filter out the noise and reduce attention to uninformative features. However, when abnormal features appear, the traditional attention mechanism is not well adapted to this goal. Thus, reevaluating the attention scores can effectively improve the accuracy and credibility of the attention mechanism [29]. The LCA shown in Fig. 3 (a) will evaluate its confidence when given an attention score, making it more solid.

For LCA, the attention branch's Q, K, and V inputs are the same: a single-modality feature. Then a regular attention score $Score_{LCA}$ and weight output $Weight_{LCA}$ are obtained after single-modality attention processing. The input of the confidence branch is the same, and the output gets a Confidence attention score $ConfScore_{LCA}$ to correct the original attention score. Due to the data of different domains having heterogeneity in content, a projection layer is responsible for mapping the domain-specific features to a common latent space. So the final V for single-modality attention is the output of the projection layer. Inspired by [16], most features are uninformative. DTMC adds a condition that conforms to a Gaussian distribution to the loss function. LCA generates the local confidence attention loss for optimizing expressed in equation 4.

$$L_{LCA} = \sum_{n=1}^{N}\sum_{l=1}^{L} \left\| \mathrm{atten}_{nl}^{sa} - \mathrm{atten}_{nl}^{lconf} \right\|_1 + \left\| \mathrm{Sk}^2 + (\text{Ku-3})^2 \right\|_1, \tag{4}$$

where $atten_{nl}^{sa}$ produced by single-modality attention and $atten_{nl}^{lconf}$ caused by local confidence evaluation. L represents the number of features. Ku and Sk represent the kurtosis and skewness of a certain distribution, respectively. If the distribution is closer to the Gaussian distribution, Ku and Sk tend to be 3 and 0.

The mean absolute error loss is used because it is less sensitive to outliers, and its penalty is fixed for any size difference. Nevertheless, this is not conducive to the convergence of the function and the learning of the model, so a lower learning rate is set.

b) Global Confidence Attention: Different modalities have different construction forms leading to composition heterogeneity. We proposed the GCA to capture multi-modality information from a global perspective to fuse multi-modality information better. The GCA is shown in Fig. 3 (b). Note that the Q input to the attention branch of GCA comes from one modality, while the K, V comes from another modality, and the attention branch outputs cross-modality attention scores and cross-modality feature weights. In addition, an M-dimensional multi-modality attention score is output, which measures the attention required for different modalities throughout the training. The input Q and K of the confidence layer also come from two modalities, and the output is a confidence score. In the same role as LCA, GCA can evaluate the cross-modality attention score. Additionally, because the modality importance varies between samples, a gated network from the sample level captures features before the fusion. Similar to LCA, global confidence attention generates the global attention confidence loss as follows:

$$L_{GCA} = \sum_{n=1}^{N}\sum_{l=1}^{L} \left\| \mathrm{atten}_{nl}^{ca} - \mathrm{atten}_{nl}^{gconf} \right\|_1 + \left\| \mathrm{Sk}^2 + (\text{Ku-3})^2 \right\|_1, \tag{5}$$

where the $atten_{nl}^{ca}$ produced by cross-modality attention and $atten_{nl}^{gconf}$ caused by global confidence evaluation. Same as equation 4, Ku and Sk represent the kurtosis and skewness of a certain distribution, respectively.

The total loss of CAM is as follows, which is the sum of the GCA and each modality's LCA:

$$L_{CA} = L_{GCA} + \sum_{m=1}^{M} L_{LCAm}. \tag{6}$$

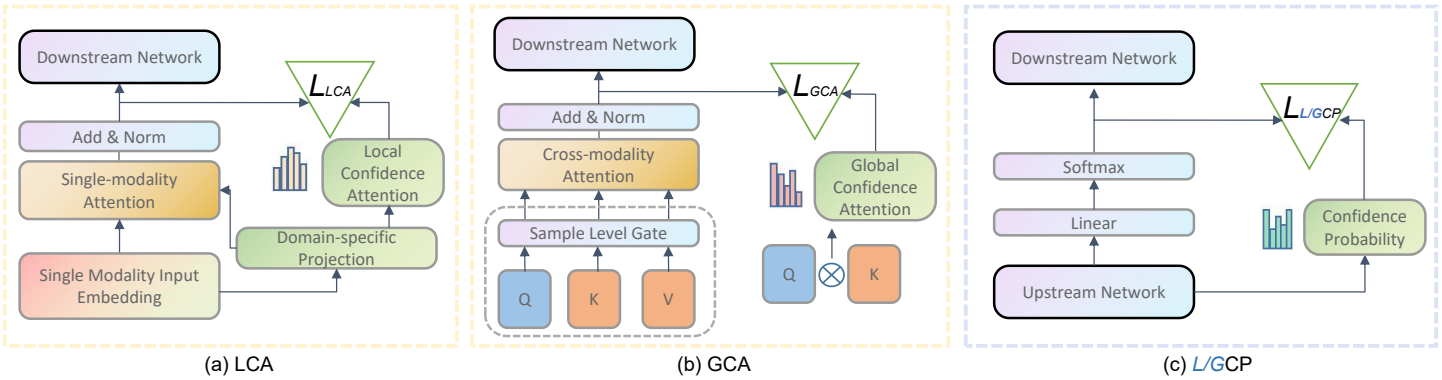

Fig. 3. Details of LCA, GCA, LCP, and GCP. Due to the same structure of LCP and GCP, we arrange them in a single figure.

D. Confidence Probability Module

The CPM includes Local Confidence Probability (LCP) and Global Confidence Probability (GCP) to improve the credibility of the results from a global and local perspective, which can be seen in Fig. 3 (c). LCP will evaluate the probability of each modality feature output by the softmax function. For C classification problems, LCP will select the corresponding prediction scores of C classes and treat them equally through the confidence layer to output a confidence score. GCP, on the other hand, focuses on evaluating the confidence of multi-modality fusion features. The different modality information contained in different samples is different, and the proportion of informativeness varies for different modalities. So using local and global attention can better capture effective information. Equations 7 and 8 are the expressions of the LCP and GCP losses.

$$L_{LCP} = \sum_{n=1}^{N}\sum_{c=1}^{C}\left\|\mathrm{prob}_{nc}^{lsof} - \mathrm{prob}_{nc}^{lconf}\right\|_1, \tag{7}$$

$$L_{GCP} = \sum_{n=1}^{N}\sum_{c=1}^{C}\left\|\mathrm{prob}_{nc}^{gsof} - \mathrm{prob}_{nc}^{gconf}\right\|_1, \tag{8}$$

where $prob_{nc}^{lsof}$ and $prob_{nc}^{lconf}$ are the raw probabilities and the confidence probabilities in single-modality features. $prob_{nc}^{gsof}$ and $prob_{nc}^{gconf}$ are the raw probabilities and the confidence probabilities in multi-modality features. The total loss of the CPM is the sum of GCP and mean of LCP in every modality, be showing as follows:

$$L_{CP} = L_{GCP} + \frac{1}{M}\sum_{m=1}^{M} L_{LCP_m}. \tag{9}$$

E. Optimization goal

The binary cross-entropy (BCE) loss is used as the final classification loss, which can be expressed as follows:

$$L_{CLS}(y, y\prime) = \sum_{n=1}^{N}\sum_{c=1}^{C} -y_{nc}\log(y_{nc}\prime) - (1-y_{nc})\log(1-y_{nc}\prime), \tag{10}$$

where y is the set of ground truth labels, and $y\prime$ is the labels predicted by the classifier and supported by CAM and CPM. Considering that some features cannot provide information but attract more attention, we define $sharp-features$ as those corresponding values with higher attention scores in the attention output layer but lower in the confidence layer. A penalty term for smoothing $sharp-features$ is introduced. Note that only the $sharp-features$ of single-modality is considered in this paper, and the $sharp-features$ of multi-modality have yet to be explored. The regularization term considerably impacts the correctness of the confidence [24] and can also help reduce the impact of $sharp-features$. Therefore, we add a penalty mechanism to the loss function as follows:

$$L_{SF} = \sum_{m=1}^{M}\sum_{r=1}^{R}\left\|W'_{mr}\right\|_1, \tag{11}$$

where M represents the number of modalities, and R is a hyperparameter indicating the number of $sharp-features$ expected to be penalized. W'_{mr} is a $sharp-feature$.

The optimization goal of DTMC is to minimize the value of equation 12, which consists of four parts in total.

$$L_{\text{OVERALL}} = L_{CLS} + \lambda_1 L_{CA} + \lambda_2 L_{CP} + \lambda_3 L_{SF}, \tag{12}$$

where L_{CLS} is the classification loss, L_{CA} is the attention loss, L_{CP} is the probability loss, and L_{SF} is the penalty term for $sharp-features$. λ_1, λ_2, λ_3 are hyperparameters that control the influence of L_{CA}, L_{CP}, and L_{SF}, respectively.

IV. Experiments and Results

This part first gives the experimental settings, then gives the results discussion and ablation experiments, and finally provides the parameter analysis and detail analysis.

TABLE II
IN FOUR BIOMEDICAL COMPUTING DATASETS, THE PRELIMINARY EXPERIMENTAL RESULTS OF DTMC COMPARED FIVE SOTA METHODS ON THREE METRICS. THE PERFORMANCE DATA OF THESE FIVE METHODS COMES FROM THEIR ORIGINAL PAPERS AND OTHER PUBLICATIONS.

Dataset	Metrics	MOGONET [32]	TMC [15]	CF [9]	GMU [10]	Dynamics [16]	DTMC (ours)	%Improv.
BRCA	ACC	82.9±1.8	84.2±0.5	81.5±0.8	80.0±3.9	87.7±0.3	**89.6±0.4**	2.2% ↑
	WeightedF1	82.5±1.7	84.4±0.9	81.5±0.9	79.8±5.8	88.0±0.5	**90.1±0.4**	2.4% ↑
	MacroF1	77.4±1.7	80.6±0.9	77.1±0.9	74.6±5.8	84.5±0.5	**88.3±0.5**	4.5% ↑
KIPAN	ACC	**99.9±0.2**	99.7±0.3	99.2±0.5	97.7±1.6	**99.9±0.2**	99.8±0.3	-0.1% ↓
	WeightedF1	**99.9±0.2**	99.7±0.3	99.2±0.5	97.6±1.7	**99.9±0.2**	**99.9±0.2**	-
	MacroF1	**99.9±0.2**	99.4±0.5	98.8±0.9	95.8±3.2	**99.9±0.3**	**99.9±0.2**	-
LGG	ACC	81.6±1.6	81.9±0.8	81.1±1.2	80.3±1.5	83.3±1.0	**83.9±0.6**	0.7% ↑
	F1	81.4±2.7	81.5±0.4	82.2±0.4	80.8±1.2	83.7±0.4	**85.6±0.7**	2.3% ↑
	AUC	84.0±2.7	87.1±0.4	88.1±0.4	88.6±1.2	88.5±0.4	**89.8±0.5**	1.4% ↑
ROSMAP	ACC	81.5±2.3	82.5±0.9	78.4±1.1	77.6±2.5	84.2±1.3	**86.0±1.0**	2.1% ↑
	F1	82.1±1.2	82.3±0.6	78.8±0.5	78.4±1.6	84.6±0.7	**86.7±0.8**	2.5% ↑
	AUC	87.4±1.2	88.5±0.6	88.0±0.5	86.9±1.6	91.2±0.7	**91.7±0.7**	0.6% ↑

TABLE III
IN TWO EDUCATION EVALUATION DATASETS, THE PRELIMINARY EXPERIMENTAL RESULTS OF DTMC COMPARED THREE SOTA METHODS ON THREE METRICS.

Dataset	Metrics	HMCAN [11]	MCAN [33]	HGLNet [34]	DTMC (ours)	%Improv.
MATH	ACC	56.0±0.6	57.1±0.5	56.7±0.6	**58.7±1.3**	2.8% ↑
	WeightedF1	52.7±0.9	55.4±0.9	56.1±0.9	**57.5±1.2**	2.5% ↑
	MacroF1	51.7±0.8	54.8±0.5	**56.7±1.0**	55.7±1.0	-1.8% ↓
ENGLISH	ACC	59.1±0.6	58.9±0.4	**60.7±1.3**	60.0±1.2	-1.2% ↓
	WeightedF1	51.8±0.9	51.7±1.0	53.5±1.0	**56.7±0.9**	6.0% ↑
	MacroF1	54.8±0.9	56.0±0.8	57.7±0.8	**61.1±1.0**	5.9% ↑

A. Experimental Settings

For multi-class datasets (BRCA, KIPAN), three evaluation metrics are ACC, WeightedF1, and MacroF1, and for two-class datasets (LGG, ROSMAP, Math, English), ACC, F1, and AUC are used. Experiments were performed on Linux (Ubuntu 20.04.1) with four Nvidia GeForce RTX 3090 GPUs and Intel(R) Xeon(R) Gold 6254 CPU @ 3.10GHz to do calculations. Experiments were repeated five times. The results shown in this paper are the average of these experiments.

There are five methods to be compared for biomedical computing: MOGONET, TMC, CF, GMU, and Dynamics. MOGONET through Graph Convolution Network and View Correlation Discovery Network to explore multi-omics relation for classification. TMC conducts decision fusion based on the confidence of different modalities. CF integrates multiple modalities by concatenating late-stage multi-modality representations. GMU establishes an intermediate multi-modality representation based on a combination of data. Dynamics dynamically evaluates the feature-level and modality-level information for integrating multi-modality.

Three SOTA methods are compared in the education evaluation domain: HMCAN, MCAN, and HGLNet. HMCAN feeds the representations into a multi-modality contextual attention network to fuse inter-modality and intra-modality relationships and design a hierarchical encoder to capture the semantics. MCAN gets features from different modalities and fuses them with a novel Co-attention mechanism. HGLNet proposes the Global Gated Attention and the Cross Residual Transformer to get representation, utilizing hierarchical information for multi-modality fusion.

B. Base experiments

a) Results and Discussions: The main experimental results for the datasets BRCA, KIPAN, LGG, and ROSMAP are shown in TABLE II. Experimental results show that DTMC achieves SOTA performance on four biomedical computing datasets. All indicators have improved, except for the ACC in the KIPAN dataset, which has decreased by 0.1%. The experimental results of datasets Math and English are shown in TABLE III. For two education evaluation datasets, DTMC also achieved excellent performance. Many metrics have improved, although the MacroF1 of the MATH dataset and the ACC of

TABLE IV
ABLATION EXPERIMENTAL RESULTS ON FOUR BIOMEDICAL COMPUTING DATASETS. w/o MEANS $without$. ACC, F1, AND AUC WERE TESTED FOR LGG AND ROSMAP. BRCA AND KIPAN TESTED ACC, WEIGHTEDF1, AND MACROF1.

Datasets	Methods	ACC	WeightedF1	MacroF1
BRCA	w/o both	87.1±0.4	87.4±0.3	83.7±0.5
	w/o CAM	87.7±0.4	87.2±0.5	86.8±0.5
	w/o CPM	88.1±0.3	89.1±0.4	87.2±0.3
	Proposed	**89.6±0.4**	**90.1±0.4**	**88.3±0.5**
KIPAN	w/o both	99.7±0.3	99.9±0.3	99.8±0.2
	w/o CAM	99.8±0.3	99.9±0.2	99.9±0.2
	w/o CPM	99.8±0.3	99.9±0.2	99.9±0.3
	Proposed	**99.8±0.3**	**99.9±0.2**	**99.9±0.2**
Datasets	Methods	ACC	F1	AUC
LGG	w/o both	80.6±0.5	81.7±0.6	85.6±0.5
	w/o CAM	81.0±0.6	82.4±0.6	86.9±0.3
	w/o CPM	82.9±0.4	83.7±0.7	87.7±0.8
	Proposed	**83.9±0.6**	**85.6±0.7**	**89.8±0.5**
ROSMAP	w/o both	82.1±0.9	82.6±1.0	89.5±0.7
	w/o CAM	84.7±0.8	82.8±0.9	90.0±0.6
	w/o CPM	85.0±0.8	84.1±0.9	91.4±0.7
	Proposed	**86.0±1.0**	**86.7±0.8**	**91.7±0.7**

TABLE V
ABLATION EXPERIMENTAL RESULTS ON TWO MULTI-CATEGORY EDUCATION EVALUATION DATASETS. w/o MEANS $without$. TEST CONDUCTED FOR ACC, WEIGHTEDF1, AND MACROF1.

Datasets	Methods	ACC	WeightedF1	MacroF1
MATH	w/o both	51.6±0.9	52.1±1.3	48.9±1.1
	w/o CAM	53.1±1.0	53.7±1.1	51.8±0.9
	w/o CPM	54.7±1.2	54.5±1.2	52.9±0.9
	Proposed	**58.7±1.3**	**57.5±1.2**	**55.7±1.0**
ENGLISH	w/o both	54.1±1.0	49.2±1.0	56.6±0.9
	w/o CAM	56.7±1.0	51.7±0.9	58.0±1.1
	w/o CPM	58.9±1.1	54.7±0.9	60.0±1.0
	Proposed	**60.0±1.2**	**56.7±0.9**	**61.1±1.0**

TABLE VI
EXPERIMENTAL RESULTS OF PARAMETER R SENSITIVITY ON FOUR BIOMEDICAL COMPUTING DATASETS.

Datasets	%.(R)	ACC	WeightedF1	MacroF1
BRCA	10%	89.5±0.6	89.7±0.3	87.8±1.0
	30%	**89.6±0.4**	**90.1±0.4**	**88.3±0.5**
	50%	88.7±0.5	89.9±0.5	88.3±0.5
KIPAN	10%	99.8±0.3	98.9±0.3	99.1±0.3
	30%	**99.8±0.3**	**99.9±0.2**	**99.9±0.2**
	50%	98.9±0.5	99.7±0.3	99.6±0.3
Datasets	%.(R)	ACC	F1	AUC
LGG	10%	81.5±0.6	84.7±0.6	87.8±1.0
	30%	**83.9±0.6**	**85.6±0.7**	**89.8±0.5**
	50%	82.7±0.5	84.9±0.5	88.3±0.5
ROSMAP	10%	84.8±0.7	85.9±0.3	90.1±0.6
	30%	**86.0±1.0**	**86.7±0.8**	89.9±0.7
	50%	85.9±0.5	84.7±0.8	**91.7±0.7**

the ENGLISH dataset have slightly decreased. The improvement of the effect is due to the advancement of the feature representation by CAM, the improvement of confidence in the decision by the CPM, and the introduction of the penalty term.

b) Ablation Study: In the ablation experiment, we tested the performance of models without CAM, CPM, and without both, respectively. The detailed results are shown in TABLE IV and TABLE V. The results of the ablation experiments verify the effectiveness of the CAM and CPM. DTMC has the best performance only when using CAM and CPM simultaneously. The experimental results also show that CAM's importance is more significant than CPM's. This may be since using single-modality attention for feature extraction and cross-modality attention for modality fusion representation is more critical because it lies upstream. A good feature representation can significantly reduce the interference of invalid information on the final classification. Another ablation experiment is to test the confidence evaluation of the classification results of the model. The experimental results can be seen in Fig. 4. Similar to [28], if the model obtains a more reliable prediction result, its polyline will fit more with the diagonal. It can be concluded from Fig. 4 that DTMC is closer to the diagonal line under the blessing of the dual-trust mechanism, which also means that the model confidence of DTMC is increased.

C. Parameter Analysis

To explore the sensitivity of DTMC to parameters, we performed a parametric analysis. The first term analyzes a set of hyperparameters λ_1, λ_2, and λ_3 controlling for loss effects. Four sets of settings for three hyperparameters were carried out in the experiment. The specific experimental results are shown in Fig. 5. It can be found that the first group of parameters (balanced weights) performs best, the last group of parameters takes second place, and the third and fourth groups rank last. Overall, the difference between different experimental results is trivial.

The second experiment investigates the effect of the parameter R in the $sharp-features$ penalty. Considering that different datasets use different numbers of features for training, the sensitivity validation to the number of R will be different. We select the 10%, 30%, and 50% of all found $sharp-features$ to be punished. The experimental results are shown in TABLE VI. Overall, selecting 30% of the $sharp-features$ for punishment is the most beneficial to DTMC. This may be due to the fact that too few $sharp-features$ do not affect the model enough, while too many $sharp-features$ may make less obvious $sharp-features$ be over-smoothed.

D. Detail Study

To further validate the effectiveness of the model and give its interpretability. We conducted additional experiments to verify

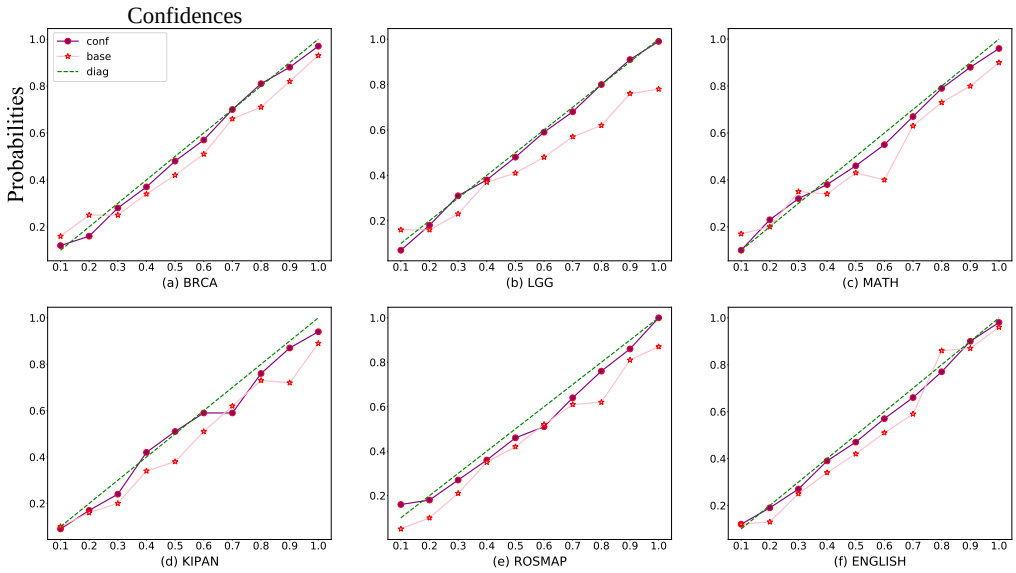

Fig. 4. Comparison of results with and without confidence mechanisms on six datasets. $conf$ indicates the result of using the confidence mechanism, $base$ indicates the result of not applying the confidence mechanism, and $diag$ indicates the diagonal.

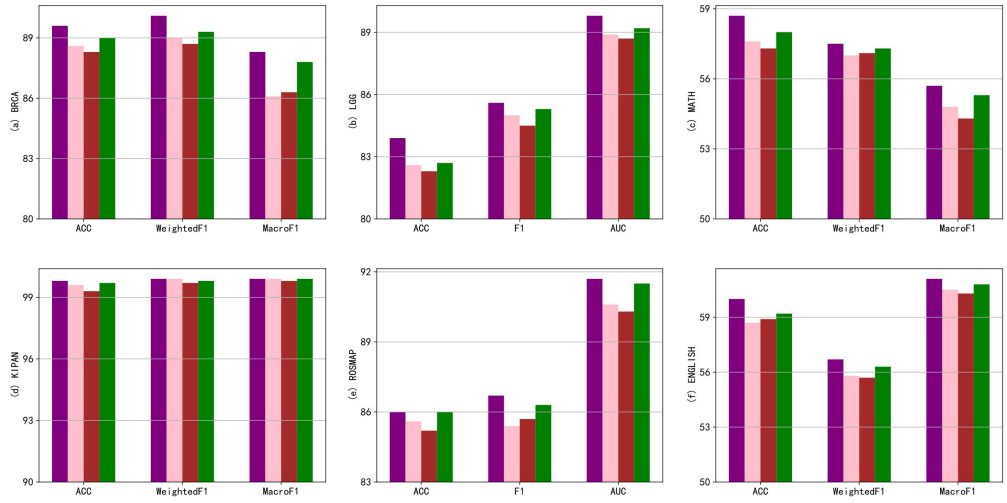

Fig. 5. Sensitivity experiment results for the parameter set λ on six datasets. A total of four sets of parameters were tested. The purple bars indicate the first set (1/3, 1/3, 1/3), the pink bars indicate the second set (1/2, 1/4, 1/4), and the brown and green bars indicate the third (1/4, 1/2, 1/4) and fourth (1/4, 1 /4, 1/2) set of parameters.

the details. The visualized results will demonstrate DTMC's performance and provide convincing reasons to reduce people's bias and distrust towards deep learning, especially in these two sensitive areas. Due to the vastness and complexity of the pre-training model, the lesson plan dataset cannot be separated into a specific discrete feature, so the penalty mechanism does not apply to this, and we have not performed parameter verification on R. This may be one of the reasons why DTMC performs slightly worse on the lesson plan dataset than on the biomedical computing dataset.

In biomedical computing, representative and essential applications of multi-modality analysis are identifying biomarkers. Some features (biomarkers) of ROSMAP and LGG with high attention scores (top 5) for each class are shown in Fig. 6 and Fig. 7. For the BRCA and KIPAN datasets, some features for obtaining high and low attention are listed in TABLE VII.

On the ROSMAP dataset, 'AD' and 'NC' represent Alzheimer's Disease patients and normal control subjects. Experiments show that biomarkers such as $SLC2A1$, $CDK18$, $SPACA6$, $hsa-miR-206$, and $CCL3$ are significant for identifying 'AD' and 'NC.' Some clinical studies such as [35], [36] have also given similar arguments. On the LGG dataset, 'Grade2' and 'Grade3' represent two grades in low-grade glioma. Experiments show that biomarkers such as $NCAPG2$, $SYT16$, $RFC2$, and $FDX1$ have high attention and significantly affect classification. The correlation between

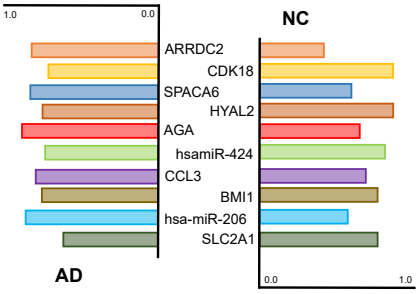

Fig. 6. The top 5 biomarkers for attention scores for each class on the ROSMAP dataset. It also shows the attention score obtained by the 5 biomarkers in another class.

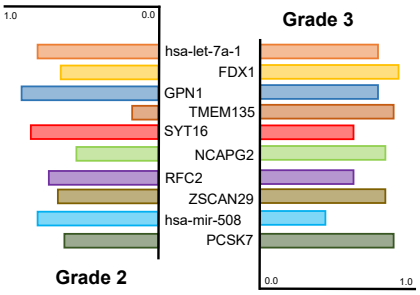

Fig. 7. The top 5 biomarkers for attention scores for each class on the LGG dataset. It also shows the attention score obtained by the 5 biomarkers in another class.

TABLE VII
THE BIOMARKERS THAT OBTAINED HIGH AND LOW ATTENTION WERE FOUND IN THE BRCA AND KIPAN DATASETS. THE PINK PART INDICATES HIGH ATTENTION, AND THE CYAN PART INDICATES LOW ATTENTION.

Datasets	Biomarkers	
	High	Low
BRCA	SOX11,KLK8,hsa-mir-452, ZNF671,TMEM207	ACOX2,BBS4,CCKBR, hsa-mir-1254,SCGB3A1
KIPAN	CHD5,ADH5,ARPC3, DVL3,NOD2	GPN1,CCDC121,hsa-mir-29b-1, TMEM232,CCDC121

these biomarkers and diseases was also found in some medical literature like [37] and [38]. The selection of these features gives the interpretability and trustworthiness of the model well, which is extremely meaningful for the safety-sensitive biomedical computing field.

We selected two lesson plans to demonstrate the effectiveness of CAM visually. The first English class is mainly about listening. The lesson plan designer made a picture with some simple objects, hoping to use the image to match the pronunciation of lesson words. The left and right half of Fig. 8 are results from self-attention and CAM, respectively, and blocks' attention scores higher than 0.8 is covered by green and red, respectively. It can be seen that regular self-attention will focus on some irrelevant but prominent features, while CAM can find relevant features more accurately. CAM can focus more on the content of this class, which is the focus of lesson plan design [39].

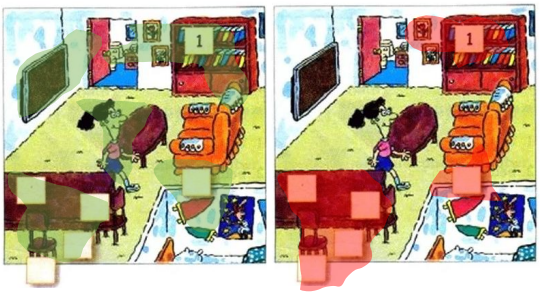

Fig. 8. The comparision of the visualization results of the image data's high-attention part, the CAM's output is on the right, and the output of the self-attention mechanism is on the left.

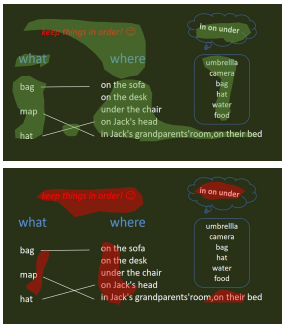

Fig. 9. Comparison of the visualization results of the high-attention part of the text data, the lower part is the output of the CAM, and the upper part is the output of the self-attention mechanism.

Fig. 9 is a visual display of some text in the second lesson plan. This class hopes to explain prepositions to students, so the core content of the lesson plan design is the knowledge of prepositions. The top half of Fig. 9 results from the self-attention mechanism, while the bottom half results from CAM. As in the previous example, using CAM will focus more on relevant features while ignoring some $sharp-features$. The above experiments give a detailed display of attention scores, which explain the final evaluation results of the model—alleviating mistrust by the public of using deep learning methods for fairness-sensitive lesson plan grading tasks.

V. CONCLUSIONS AND FUTURE WORK

In this work, we proposed DTMC for the problem of untrustworthy and uninterpretable multi-modality learning with DNN models. Extensive experimental results on four biomedical computing datasets prove that DTMC reaches the SOTA level and alleviates the lack of interpretability and confidence in previous DL methods. A penalty trick reduces the influence of $sharp-features$. Additionally, we collected and curated a multi-modality lesson plan grading dataset for testing DTMC and community use. DTMC also reaches SOTA performance on this dataset. In the future, we hope to explore the internal logic of deep neural networks from a more fine-grained perspective.

References

[1] N. Audebert, C. Herold, K. Slimani, and C. Vidal, "Multimodal deep networks for text and image-based document classification," in *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I*. Springer, 2020, pp. 427–443.

[2] H. Lei and N. Chen, "Audio-visual scene classification based on multimodal graph fusion," *Proc. Interspeech 2022*, pp. 4157–4161, 2022.

[3] C. Song, N. Ning, Y. Zhang, and B. Wu, "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks," *Information Processing & Management*, vol. 58, no. 1, p. 102437, 2021.

[4] Q. Zhong, Q. Wang, and J. Liu, "Combining knowledge and multimodal fusion for meme classification," in *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part I*. Springer, 2022, pp. 599–611.

[5] J. J. Bird, D. R. Faria, C. Premebida, A. Ekárt, and G. Vogiatzis, "Look and listen: A multi-modality late fusion approach to scene classification for autonomous machines," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 380–10 385.

[6] T. Saha, A. Patra, S. Saha, and P. Bhattacharyya, "Towards emotion-aided multi-modal dialogue act classification," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4361–4372.

[7] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," *arXiv preprint arXiv:1909.02950*, 2019.

[8] D. Kumar, N. Kumar, and S. Mishra, "Quarc: Quaternion multi-modal fusion architecture for hate speech classification," in *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2021, pp. 346–349.

[9] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multi-modal learning better than single (provably)," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 944–10 956, 2021.

[10] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.

[11] Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, "Multimodal fusion with co-attention networks for fake news detection," in *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, 2021, pp. 2560–2569.

[12] R. You, Z. Guo, L. Cui, X. Long, Y. Bao, and S. Wen, "Cross-modality attention with semantic graph embedding for multi-label classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 709–12 716.

[13] Z. Zhang, Z. Wang, X. Li, N. Liu, B. Guo, and Z. Yu, "Modalnet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network," *World Wide Web*, vol. 24, pp. 1957–1974, 2021.

[14] A. Tonge and C. Caragea, "Dynamic deep multi-modal fusion for image privacy prediction," in *The World Wide Web Conference*, 2019, pp. 1829–1840.

[15] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification," *arXiv preprint arXiv:2102.02051*, 2021.

[16] Z. Han, F. Yang, J. Huang, C. Zhang, and J. Yao, "Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 707–20 717.

[17] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, and P. Pérez, "Addressing failure prediction by learning model confidence," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[18] T. Zahavy, A. Magnani, A. Krishnan, and S. Mannor, "Is a picture worth a thousand words? a deep multi-modal fusion architecture for product classification in e-commerce," *arXiv preprint arXiv:1611.09534*, 2016.

[19] I. Gallo, A. Calefati, S. Nawaz, and M. K. Janjua, "Image and encoded text fusion for multi-modal classification," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2018, pp. 1–7.

[20] D. Kiela, E. Grave, A. Joulin, and T. Mikolov, "Efficient large-scale multi-modal classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[21] A. Mittal, K. Dahiya, S. Malani, J. Ramaswamy, S. Kuruvilla, J. Ajmera, K.-h. Chang, S. Agarwal, P. Kar, and M. Varma, "Multi-modal extreme classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 393–12 402.

[22] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang *et al.*, "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, 2021.

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

[24] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[25] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi, "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability," *Computer Science Review*, vol. 37, p. 100270, 2020.

[26] A. F. Psaros, X. Meng, Z. Zou, L. Guo, and G. E. Karniadakis, "Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons," *Journal of Computational Physics*, p. 111902, 2023.

[27] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *International conference on machine learning*. PMLR, 2020, pp. 9690–9700.

[28] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[29] Q. Li, D. Qiu, Y. Zhang, B. Li, Y. He, P. C. Woodland, L. Cao, and T. Strohman, "Confidence estimation for attention-based sequence-to-sequence models for speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6388–6392.

[30] A. Kastanos, A. Ragni, and M. J. Gales, "Confidence estimation for black box automatic speech recognition systems using lattice recurrent neural networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6329–6333.

[31] M. Rizzo, C. Conati, D. Jang, and H. Hu, "Evaluating the faithfulness of saliency-based explanations for deep learning models for temporal colour constancy," *arXiv preprint arXiv:2211.07982*, 2022.

[32] T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding, and K. Huang, "Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification," *Nature communications*, vol. 12, no. 1, p. 3445, 2021.

[33] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2021, pp. 153–162.

[34] J. Wu, J. Zhao, and J. Xu, "Hglnet: A generic hierarchical global-local feature fusion network for multi-modal classification," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.

[35] H. Zheng, G. Long, Y. Zheng, X. Yang, W. Cai, S. He, X. Qin, and H. Liao, "Glycolysis-related slc2a1 is a potential pan-cancer biomarker for prognosis and immunotherapy," *Cancers*, vol. 14, no. 21, p. 5344, 2022.

[36] W. Zhang, J. I. Young, L. Gomez, M. A. Schmidt, D. Lukacsovich, A. Varma, X. S. Chen, E. R. Martin, and L. Wang, "Distinct csf biomarker-associated dna methylation in alzheimer's disease and cognitively normal subjects," *Alzheimer's Research & Therapy*, vol. 15, no. 1, p. 78, 2023.

[37] W. Huang, Y. Wu, J. Zhu, N. Luo, C. Wang, S. Liu, and Z. Cheng, "Pan-cancer integrated bioinformatics analysis reveals cuproptosis related gene fdx1 is a potential prognostic and immunotherapeutic biomarker for lower-grade gliomas," *Frontiers in Molecular Biosciences*, vol. 10, 2023.

[38] J. Chen, Z. Wang, W. Wang, S. Ren, J. Xue, L. Zhong, T. Jiang, H. Wei, and C. Zhang, "Syt16 is a prognostic biomarker and correlated with immune infiltrates in glioma: a study based on tcga data," *International Immunopharmacology*, vol. 84, p. 106490, 2020.

[39] M. of Education of the People's Republic of China, "General senior high school english curriculum standards (2017 edition, 2020 revision)," 2020.