Multimodal Dynamics: Dynamical Fusion for Trustworthy Multimodal Classification

Zongbo Han[1*‡], Fan Yang[2*], Junzhou Huang[2], Changqing Zhang[1†], Jianhua Yao[2†]
College of Intelligence and Computing, Tianjin University[1],
Tencent AI Lab[2]
{zongbo,zhangchangqing}@tju.edu.cn, {fionafyang,jianhuayao}@tencent.com, jzhuang75@gmail.com

Abstract

Integration of heterogeneous and high-dimensional data (e.g., multiomics) is becoming increasingly important. Existing multimodal classification algorithms mainly focus on improving performance by exploiting the complementarity from different modalities. However, conventional approaches are basically weak in providing trustworthy multimodal fusion, especially for safety-critical applications (e.g., medical diagnosis). For this issue, we propose a novel trustworthy multimodal classification algorithm termed Multimodal Dynamics, which dynamically evaluates both the feature-level and modality-level informativeness for different samples and thus trustworthily integrates multiple modalities. Specifically, 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. Then a transparent fusion algorithm based on the dynamical informativeness estimation strategy is induced. To the best of our knowledge, this is the first work to jointly model both feature and modality variation for different samples to provide trustworthy fusion in multi-modal classification. Extensive experiments are conducted on multimodal medical classification datasets. In these experiments, superior performance and trustworthiness of our algorithm are clearly validated compared to the state-of-the-art methods.

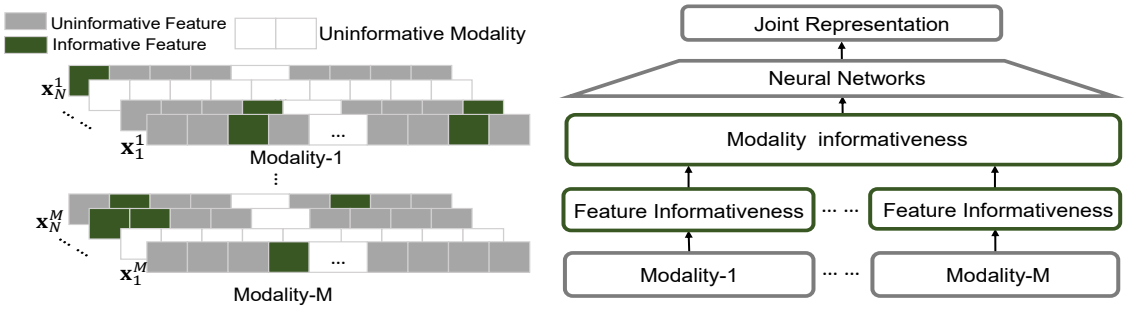
1. Introduction

Multimodal learning has achieved impressive success in a wide spectrum of applications (e.g., medical-diagnosis [16, 52]), which improves the performance by exploring the complementary information from different modalities. Representative multimodal methods typically integrate different modalities into a unified representation with powerful neural networks [29, 34, 45, 61, 63, 64, 71, 72, 74]. Despite encouraging progress, traditional multimodal models are still unreliable due to the limitation of existing fusion strategies. As a result, existing multimodal learning also challenges itself in deployment for safety-critical applications (e.g., computer-aided diagnosis). This inspires us to utilize multimodal information in a more elegant way to produce trustworthy multimodal fusion.

For multimodal learning, traditional methods mainly focus on obtaining a common or joint representation by exploring the correlated and complementary information between different modalities with powerful neural networks [8, 65]. Some existing multimodal methods obtain a joint representation by simply concatenating the features obtained from different modalities [26, 32]. Then a neural network is employed to explore the joint representation. Besides, joint representations can be obtained through carefully designed objective functions [3, 4, 27, 63] and neural network architectures [6, 38, 62]. Although effective, these methods are weak in dynamically perceiving the informativeness of each feature and modality for different samples, which could enhance the trustworthiness (including stability and explainablity) in multimodal classification. In multimodal medical data, as shown in Fig. 1, uninformative features and modalities widely exist due to the unsatisfactory data collection (e.g., inherent noise in multiomics data [7], uneven quality of histopathological images for different patients [68] and tabular data with complex missing patterns and feature noise [70]). This motivates us to evaluate the informativeness of each feature and each modality of different samples, and conduct a dynamical multimodal fusion.

In this work, we propose a novel algorithm termed *Multimodal Dynamics* for trustworthy multimodal classification, which models the feature and modality informativeness to promote the fusion stability and explainablity. Specifically, we introduce a sparse gating strategy to dynamically obtain the informative features for different samples, and the modality confidence is introduced to dynamically evalu-

∗ Equal contribution. † Corresponding authors. ‡ Supported by 2021 Tencent Rhino-Bird Research Elite Training Program.



(a) Dynamics in multimodal data.

(b) Proposed multimodal learning paradigm.

Figure 1. (a) Illustration of feature and modality dynamics in multimodal data. For one modality, the informativeness of different features may vary with the samples. Meanwhile, the informativeness of different modalities may also change for different samples. (b) To capture the dynamics, multimodal dynamics paradigm is proposed, where feature and modality informativeness is dynamically evaluated to promote multimodal fusion.

ate the informativeness of different modalities for different samples. Accordingly, a unified multimodal fusion framework is introduced to dynamically fuse informative features and modalities, and to reduce the influence from noisy features and modalities, endowing the model with robustness for dynamic variation of quality for features and modalities, and trustworthiness for final decision. For clarification, the contributions of our method could be summarized as follows: (i) We propose a dynamical multimodal fusion strategy, which models both the feature-level and modality-level dynamicities to provide a trustworthy multimodal fusion. To the best of our knowledge, the proposed method is the first work to exploit the feature-level and modality-level dynamicities for trustworthy multimodal fusion. (ii) We introduce effective mechanisms, i.e., sparse gating and true class probability approximation to dynamically estimate the dynamicity of each feature and modality, which are cooperative for the optimal prediction. (iii) We conduct experiments on four multimodal medical classification datasets and the experimental results demonstrate significant improvement against state-of-the-art methods. Qualitative experiments also validate the trustworthiness and interpretability in modeling the multimodal dynamicity. [1]

2. Related Work

Multimodal learning. To integrate multiple types of data for decision making, multimodal learning has been widely explored recently [8, 48, 65]. Existing multimodal methods are typically divided into early [47], intermediate [6, 26, 29, 30, 32, 33, 37, 38, 58, 66] and decision [23,43,53,55,62] fusion according to the fusion strategies [8, 48]. Early fusion based methods directly integrate multiple modalities at the data level, typically concatenating multimodal data [47], which may fail to handle high-dimensional or heterogeneous data. Intermediate fusion strategy is widely adopted in multimodal learning, which allows multiple modalities to be fused at any layer through a well-designed network [6,26,29,30,32,33,37,38,58,66]. For some methods, the intermediate representations from different modalities are concatenated to obtain a joint representation [26, 32]. Gated multimodal fusion [6] aims to find an intermediate multimodal representation based on the combination of features from different modalities. Besides, decision fusion can perform multimodal fusion based on the uncertainty of prediction [23, 39, 55, 56]. There have been methods that focus on the dynamics between different modalities [46, 57]. Note that, none of the above methods pay attention to the dynamics of the features and modalities simultaneously for trustworthy classification.

Uncertainty learning. Although deep learning has achieved great success in many applications, it is hard to provide reliable predictive uncertainty or confidence [2, 19, 21], which is crucial for trustworthy models. Bayesian methods [9,44,67] provide the predictive uncertainty by replacing the deterministic parameters with the distribution. However, the computationally intensive nature of Bayesian methods limits the applicability in deep neural networks. MC-dropout [20] applies dropout at both training and test stages to avoid the computational cost. Ensemble-based methods [5, 24, 36] train and integrate multiple deterministic neural networks to calculate the predictive uncertainty. Different from the uncertainty estimation algorithms, confidence calibration methods [11, 22, 51] aim to obtain confidence by calibrating the classification results directly. In this paper, we employ a confidence based model to assess the informativeness of different modalities for each sample.

3. Proposed Method

In this section we elaborate the proposed multimodal classification algorithm. Given N i.i.d. multimodal ob-

[1] Code is available at github.com/TencentAILabHealthcare/mmdynamics.

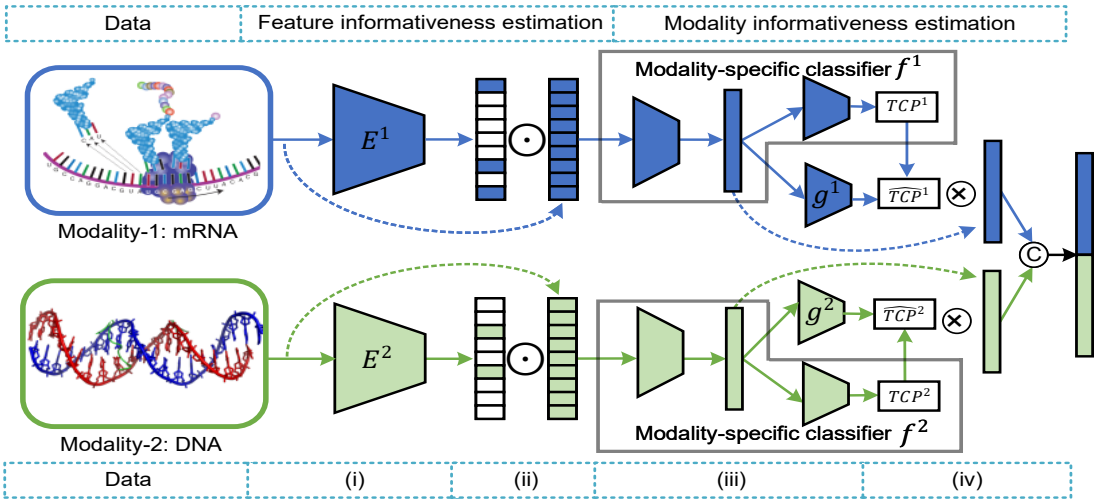


Figure 2. Framework of multimodal dynamics. We use a two-modality case for better illustration. The proposed method is mainly composed of the following steps. (i) For modality m, the sparse feature informativeness is obtained with encoder E^m. (ii) A gating strategy is employed to preserve the informative features, where the dotted line with arrow indicates the move of original data. (iii) A confidence regression network g^m is used to approximate TCP, which is the predictive probability of modality-specific classifier f^m corresponding to the real label. The obtained $\widehat{TCP}$ reflects the informativeness of different modalities. (iv) A gated network is introduced at modality-level to dynamically fuse multiple modalities based on the informativeness.

servations with M modalities and the corresponding labels $\{\{\mathbf{x}_n^m\}_{m=1}^M, \mathbf{y}_n\}_{n=1}^N$, the goal of multimodal classification is to construct a mapping between multimodal data $\{\mathbf{x}_n^m \in \mathbb{R}^{d_m}\}_{m=1}^M$ and the class label $\mathbf{y}_n \in \mathbb{R}^K$, where d_m and K are the dimensionality of feature space for the m modalities and the number of classes, respectively. Formally, to integrate multimodal information and learn the underlying mapping between the multimodal observations and the class labels, a neural network $f : \{\mathbf{x}^m\}_{m=1}^M \rightarrow \mathbf{y}$ is trained in conventional multimodal classification algorithms. To achieve a more trustworthy integration, unlike the previous algorithms, the proposed multimodal classification algorithm models both the feature-level (elaborated in detail in Section 3.1) and modality-level dynamics (elaborated in detail in Section 3.2). Then a dynamical multimodal fusion algorithm is proposed in Section 3.3.

3.1. Feature-level Dynamics

Given a high-dimensional feature vector $\mathbf{x}^m \in \mathbb{R}^{d_m}$, there is usually a subset of features relevant to the class label, reflecting the informativeness of different features in classification [15, 17]. Accordingly, sparsity induced models are popular in handling high-dimensional data. Differently, we argue that the informativeness of different features are different and more importantly, the informativeness for one feature is dynamically changed for different samples, which should be considered during the multimodal fusion. By modeling the dynamic, our algorithm is endowed with the following merits: (i) retaining important features and removing redundant and noisy ones, thereby promoting the multimodal fusion; (ii) enhancing the explanation ability of the multimodal fusion. To this end, we introduce a dynamical feature informativeness coding network to retain the informative features and suppress the uninformative features in different modalities, which stabilizes and promotes the within-modality representation.

Feature-informativeness encoder. To identify the feature-level informativeness, we train an encoder network $E^m : \mathbf{x}^m \rightarrow \mathbf{w}^m$, where $\mathbf{w}^m \in \mathbb{R}^{d_m}$ refers the feature informativeness vector. Besides, to obtain a more intuitive informativeness vector, sigmoid activation is used, which could allow the output of the E^m to be scaled:

$$\mathbf{w}^m = \sigma(E^m(\mathbf{x}^m)) = [w_1^m, \cdots, w_{d_m}^m], \tag{1}$$

where σ refers to the sigmoid activation function. Accordingly, the dynamics of features for different samples are modeled. For high-dimensional data, we incorporate the sparsity prior which seeks a small subset of relevant features. Specifically, to promote the sparsity, ℓ_0 regularization is employed:

$$\mathcal{L}_{\ell_0}^s = \sum_{m=1}^{M}\sum_{d=1}^{d_m} s_d^m, \text{ with } s_d^m = \begin{cases} 1 & \text{if } \ w_d^m \neq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

Since ℓ_0-norm is hard to optimize in practice, ℓ_1-norm is adopted for approximation:$\mathcal{L}_{\ell_1}^s = \sum_{m=1}^{M} \|\mathbf{w}^m\|_1$, where $\|\mathbf{w}\|_1$ represents the ℓ_1-norm of $\mathbf{w}$. Intrinsically, we introduce a dynamic sparsity strategy in our algorithm.

3.2. Modality-level Dynamics

For multimodal data, the informativeness of a modality is basically not fixed for different samples [28, 49]. Therefore, it is crucial for multimodal classification to be aware of the informativeness variation with respect to different samples, which is related to whether the model can adapt to the quality variation of modalities. Based on the above motivation, we employ the True-Class-Probability [11] to quantify the classification confidence of different modalities, which is closely related to the modality informativeness for classification. When the confidence of a modality classification is low, it means that the classification is uncertain, and the informativeness of the corresponding modality is low and vice versa.

Maximum class probability. To obtain the classification confidence of different modalities, M classifiers $f^m : \mathbf{x}_n^m \to \mathbf{y}_n$ are constructed. For modality m, a classification neural network f^m can be regarded as a probabilistic model, which converts an observation $\mathbf{x}^m$ to a predictive distribution $\mathbf{p}^m(\mathbf{y} \mid \mathbf{x}^m) = [p_1^m, \cdots, p_K^m]$ based on the Softmax output. The classifier can be trained with a maximum likelihood estimation framework to minimize the Kullback-Leibler divergence between the predictive distribution and the true distribution:

$$\mathcal{L}^{cls} = -\sum_{m=1}^{M}\sum_{k=1}^{K} y_k \log p_k^m, \tag{3}$$

where y_{ik} is the k-th element of the class label $\mathbf{y}_i$. Eq. 3 is also known as cross-entropy loss function. Then the maximum class probability can be inferred with $MCP^m = \max\{p_1^m, \cdots, p_K^m\}$, which can be considered as the confidence of the classifier for the prediction.

Multimodal confidence. Although effective in classification, MCP usually leads to over-confidence especially for erroneous prediction [41, 59]. Therefore, the true-class-probability (TCP) is employed to obtain more reliable classification confidence. Different from MCP, which uses the largest Softmax outputs as confidence, TCP uses the Softmax output probability corresponding to the real label as the confidence. Formally, for modality m, given the prediction distribution $\mathbf{p}^m(\mathbf{y} \mid \mathbf{x}^m) = [p_1^m, \cdots, p_K^m]$ and the corresponding label $\mathbf{y}$, TCP^m can be written as

$$TCP^m = \mathbf{y} \cdot \mathbf{p}^m(\mathbf{y} \mid \mathbf{x}^m) = \sum_{k=1}^{K} y_k p_k^m, \tag{4}$$

where $(\cdot)$ defines the inner product. It is easy to understand that for correctly classified samples, TCP is equivalent to MCP. At this time, TCP and MCP are both the largest Softmax outputs, which could promisingly reflect the classification confidence. However, when misclassified, TCP can better reflect the classification than MCP because TCP would be more likely to be close to a low value, reflecting the fact that the model tends to make an erroneous prediction.

Although TCP can obtain more reliable confidence, it cannot be used in the test stage directly due to the need of label information. Therefore, for modality m, a confidence neural network $g^m : \mathbf{x}^m \to TCP^m$ is introduced to approximate TCP^m. Since the $TCP \in (0, 1)$, sigmoid activate function is employed in the last layer of the neural network and ℓ_2 loss is used to train the confidence neural networks:

$$\mathcal{L}^{conf} = \sum_{m=1}^{M} (\widehat{TCP}^m - TCP^m)^2 + \mathcal{L}^{cls}, \tag{5}$$

where $\widehat{TCP}^m = g^m(\mathbf{x}^m)$. Then the TCP can be approximated with the modality-specific classifier and confidence regression network.

3.3. Dynamical Multimodal Fusion

According to Section 3.1 and Section 3.2, feature-level informativeness $\{\mathbf{w}^m\}_{m=1}^M$ and modality-level informativeness $\{\widehat{TCP}^m\}_{m=1}^M$ can be obtained respectively. In this section, we elaborate how to conduct dynamical multimodal fusion based on the feature and modality informativeness. To achieve this goal, a nested fusion structure is considered. The framework of the model can be referred to Fig. 2.

Firstly, we consider the feature informativeness in classification. Given a feature vector $\mathbf{x}^m \in \mathbb{R}^{d_m}$, the feature informativeness vector $\mathbf{w}^m$ can be obtained with $\mathbf{w}^m = \sigma(E^m(\mathbf{x}^m))$. Then a gating strategy is used to incorporate the informativeness information of features, which could allow the informative features to be retained and enforce the uninformative features to be suppressed: $\widetilde{\mathbf{x}}^m = \mathbf{x}^m \odot \mathbf{w}^m$, where $\odot$ represents the element-wise multiplication.

Secondly, we consider the modality informativeness in classification. According to Section 3.2, modality specific classifier f^m and confidence regression network g^m are trained to estimate the classification confidence. To make use of the information of each modality-specific classifier f^m, we use f_1^m to extract the information of each modality where f_1^m is f^m with the last fully connected layer removed. Formally, we can obtain the late representation of each modality with $\mathbf{h}^m = f_1^m(\widetilde{\mathbf{x}}^m)$. Meanwhile, the modality confidence can be estimated with $\widehat{TCP}^m = g^m(\mathbf{h}^m)$. A modality-level gating strategy is employed to incorporate the modality informativeness:

$$\mathbf{h} = [\widehat{TCP}^1 \mathbf{h}^1, \cdots, \widehat{TCP}^M \mathbf{h}^M], \tag{6}$$

where $[\cdot, \cdot]$ is the concatenation operator and $\mathbf{h}$ is the multimodal representation. An additional classifier $f : \mathbf{h} \to \mathbf{y}$ is trained with cross-entropy loss $\mathcal{L}^f$ to obtain

the final multimodal classification results $\mathbf{p}$, where $\mathcal{L}^f = -\sum_{k=1}^{K} y_k \log p_k$. The overall loss function can be written as

$$\mathcal{L} = \sum_{i=1}^{N} (\mathcal{L}^f + \lambda_1 \mathcal{L}^s_{\ell_1} + \lambda_2 \mathcal{L}^{conf}), \tag{7}$$

where λ_1 and λ_2 are hyperparameters used to balance different losses. The model can be obtained by optimizing the loss $\mathcal{L}$.

4. Experiments

In the experiment, we compare the proposed method with current state-of-the-art classification algorithms on four real-world datasets. The extensive experimental results clearly illustrate the superiority of the proposed method. In addition, we also conduct ablation study which indicates that the proposed dynamical fusion indeed promotes multimodal classification.

4.1. Experimental Setup

Datasets. We conduct extensive experiments on four real-world multimodal medical datasets. **BRCA** for breast invasive carcinoma PAM50 subtype classification contains 875 samples of 5 different classes. **LGG** for grade classification in glioma contains 510 samples of 2 classes. **ROSMAP** for Alzheimer's Disease diagnosis is composed of ROS [1] and MAP [14], which contains 351 samples of 2 classes. **KIPAN** for kidney cancer type classification contains 658 samples of 3 classes. The above datasets are associated with three different modalities including mRNA expression data, DNA methylation data, and miRNA expression data. BCRA, LGG, and KIPAN can be acquired from The Cancer Genome Atlas program (TCGA)[2].

Compared methods. To investigate the improvement of the multimodal fusion strategy, we compare our method with 5 single-modal classification methods trained with the simple concatenation of the multimodal data (early fusion), including K-Nearest Neighbors (**KNN**) [18], Support Vector Machine (**SVM**) [12], Linear Regression (**LR**) trained with ℓ_1 regularization, Random Forest classifier (**RF**) [25], and fully connected neural networks (**NN**). We also compare our method with currently 7 state-of-the-art multimodal classification models. Group-regularized (logistic) ridge regression (**GRidge**) [60] makes structural use of multimodal data through group-specific penalties. Block partial least squares discriminant analysis (**BPLSDA**) [54] explores multimodal data in latent space through discriminant analysis. Block sparse partial least squares discriminant analysis (**BSPLSDA**) [54] selects the most relevant features by adding sparse constraints to BPLSDA. Multiomics graph convolutional networks (**MOGONET**) [62] constructs a graph for every modalities and explores multimodal correlation via correlation discovery network. Concatenation of final multimodal representations (**CF**) [26,32] integrates multiple modalities by concatenating late stage multimodal representations. Gated multimodal units for information fusion (**GMU**) [6] establishes an intermediate multimodal representation based on a combination of data. Trusted multiview classification (**TMC**) [23] conducts decision fusion based on the confidence of different modalities.

Evaluation metrics and experimental details. We employ three metrics, i.e., accuracy (ACC), F1 score (F1), and area under the receiver operating characteristic curve (AUC), to evaluate the performance of different methods for binary classification tasks. For the multi-class classification tasks, we report the experiment results in terms of accuracy (ACC), average F1 score, average F1 score weighted by support (WeightedF1), and macro-averaged F1 score (MacroF1). We employ the same experimental settings as in [62]. We run experiments 20 times to report the mean and standard deviation. The Adam optimizer [35] with learning rate decay is employed to train the model.

4.2. Quantitative Analysis

Multi-class classification. Firstly, we compare the proposed methods with state-of-the-art single modal and multimodal classification methods on multi-classification tasks. The detailed experimental results on BRCA and KIPAN are shown in Table 1. The following conclusions can be drawn from the experimental results. (i) The proposed method outperforms other methods on most datasets. Taking the results on BRCA for example, the ACC of the proposed method is 87.7% while the second best methods (TMC) is 84.2%. (ii) Benefiting from exploring the multimodal information, the proposed method is consistently better than the single modal algorithms on all datasets. For example, on the BRCA dataset, our proposed method achieves significant improvement around 12.3%, 14% and 17.7% over the most competitive method in terms of ACC, WeightedF1 and MacroF1, respectively. (iii) Compared with other multimodal algorithms, the proposed method has a significant performance improvement on most datasets. Intuitively, the possible reason is that the proposed method reduces the irrelevant information through dynamical fusion.

Binary classification. We further conduct comparison experiments on the binary classification task. Table 2 demonstrates the classification results on LGG and ROSMAP in terms of ACC, F1, and AUC respectively. The proposed method achieves the best performance compared with the other methods in terms of ACC and F1. The proposed algorithm achieves 1.7% and 2.3% improvements over the second performer TMC in terms of ACC and F1. Our multimodal dynamics outperforms the single-modal classification methods thanks to the flexible and effective

[2]https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

		BRCA			KIPAN		
Method	Fusion strategy	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
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
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
Ours	dynamical	**87.7±0.3**	**88.0±0.5**	**84.5±0.5**	**99.9±0.2**	**99.9±0.2**	**99.9±0.3**

Table 1. Comparison with state-of-the-art methods on the BRCA and KIPAN datasets, where the best results are in bold.

		LGG			ROSMAP		
Method	Fusion strategy	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
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
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
Ours	dynamical	**83.3±1.0**	**83.7±0.4**	88.5±0.4	**84.2±1.3**	**84.6±0.7**	**91.2±0.7**

Table 2. Comparison with state-of-the-art methods on the LGG and ROSMAP datasets, where the best results are in bold.

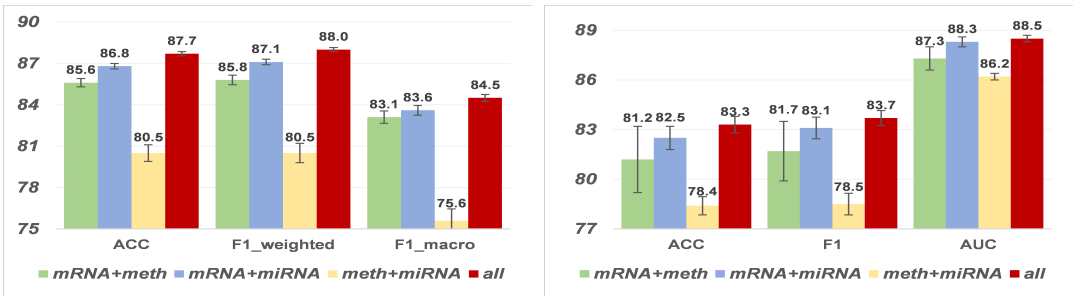
multimodal fusion. For example, there are at least 7.2%, 6.8%, and 8.5% improvements over the best single-modal classification methods in terms of ACC, F1, and AUC respectively.

Ablation study. We further perform ablation study on these four datasets. Specifically, we compare the proposed methods with concatenation of final multimodal representations (CF), sparse feature informativeness induced integration (FI), and modality informativeness induced integration (MI). Table 3 provides the results of the ablation study. According to the results, we have the following observations. (i) Both the FI and MI outperform the simple concatenation of final multimodal representations. The possible reason is that the obtained informativeness could dynamically guide the multimodal fusion. (ii) Benefiting from the more comprehensive informativeness information during fusion, the dynamical multimodal fusion produces more promising results on most datasets. For example, our method achieves at least 0.8% improvement in terms of ACC on BRCA dataset (87.7% vs 86.9, p-value of t-test: 0.00055).

Performance of the proposed method under different modalities data types. To demonstrate the necessity of integrating multiple modalities data, we compare the different settings with different combinations of the three available modalities including mRNA+meth, mRNA+miRNA, and meth+miRNA on BRCA and LGG datasets, where mRNA, meth, and miRNA refer to the mRNA expression, DNA methylation and miRNA expression data respectively. The experimental results are shown in Fig. 3. Benefiting from the integration of more comprehensive information, the method using three different modalities achieves the best performance.

4.3. Qualitative Analysis

We further conduct qualitative analysis to intuitively investigate the superiority and effectiveness of the introduced feature informativeness and modality informativeness modules. Specifically, the following experiments are conducted:

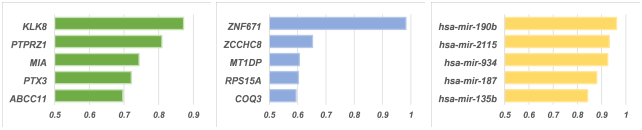


(a) BRCA (b) LGG

Figure 3. Performance comparison of different modalities classification via the proposed methods on BRCA and LGG datasets.

Dataset	Method	ACC	WeightedF1	MacroF1
BRCA	CF	81.5±0.8	81.5±0.8	77.1±0.9
	FI	84.9±0.6	85.1±0.6	81.9±0.7
	MI	86.9±0.9	87.2±0.9	83.2±0.9
	Proposed	**87.7±0.3**	**88.0±0.3**	**84.5±0.5**
KIPAN	CF	99.2±0.5	99.2±0.5	98.8±0.9
	FI	99.8±0.2	99.8±0.2	99.9±0.3
	MI	99.5±0.3	99.5±0.3	99.4±0.5
	Proposed	**99.9±0.2**	**99.9±0.2**	**99.9±0.3**
Dataset	Method	ACC	F1	AUC
LGG	CF	81.1±1.2	82.2±1.0	88.1±0.4
	FI	82.4±1.4	82.6±1.4	88.3±0.6
	MI	82.9±0.8	83.1±0.7	**90.2±0.2**
	Proposed	**83.3±1.0**	**83.7±0.9**	88.5±0.4
ROSMAP	CF	78.4±1.1	78.8±0.9	88.0±0.5
	FI	80.4±1.9	81.3±1.7	88.4±1.3
	MI	83.8±1.3	84.2±1.2	90.7±0.9
	Proposed	**84.2±1.3**	**84.6±1.2**	**91.2±0.7**

Table 3. Ablation study on the BRCA, KIPAN, LGG and ROSMAP datasets, where the best results are in bold. For clarity, LF, FI and MI in the table indicate the simple late fusion, sparse feature informativeness induced fusion and modality informativeness induced fusion respectively.



(a) mRNA expression (b) DNA methylation (c) miRNA expression

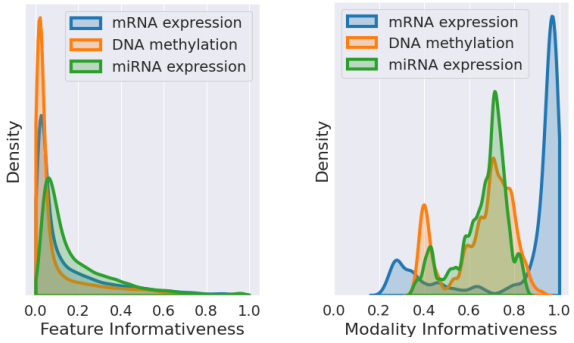
Figure 4. The top 5 informative biomarkers from different modalities on BRCA dataset identified by our algorithm.

(i) biomarkers identification via the obtained feature informativeness; (ii) density estimation of the obtained feature informativeness to illustrate the effect of the employed ℓ_1 loss; (iii) density estimation of the obtained modality informativeness to illustrate the effect of the employed modality informativeness strategy; (iv) visualization of the obtained feature and modality informativeness with heatmap.

Biomarkers identification. The representative and important application of multiomics analysis is to identify biomarkers for early diagnosis and prognosis, and to discover drug targets for treatments. To this end, we investigate biomarker identification and drug target discovery by interpreting the feature informativeness of the multiomics data in our Multimodal Dynamics. Specifically, due to the randomness involved, we run experiments 5 times to obtain the mean of the feature informativeness of all samples on the test samples. Note that although the results shown are the mean of all samples for ease of explanation, our algorithm could provide the feature informativeness for each sample.

Taking the BRCA data set as an example, the top 5 important features of the three modalities are shown in Fig. 4 and their participation in breast cancer progression and indication effects can be verified through reviewing previous biological and medical studies. Here we briefly introduce some representative researches. KLK8 is downregulated in breast cancer and has been verified as an independent indicator of the prognosis of breast cancer patients [40]. The change of PTPRZ1's expression is reported to promote tumor proliferation and inhibit apoptosis in breast cancer cells [31]. Elevated levels of MIA protein are detected in the serum of patients with advanced-stage breast cancers [10]. PTX3 shows significantly higher expression in breast-infiltrating carcinomas [50]. ABCC11 is highly expressed in aggressive breast cancer subtypes, and tumor ABCC11 expression is associated with poor prognosis [69]. ZNF671 plays a tumor suppressor role in breast cancer [73]. The miR-190b regulates cell progression and acts as potential biomarkers for breast cancer [13] and the miR-187 is an independent prognostic factor in breast cancer [42].

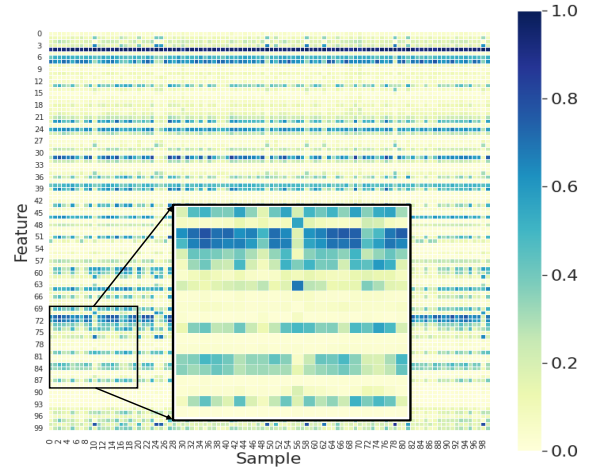
Density of feature informativeness. To visualize the obtained feature informativeness, a kernel density estimate (KDE) plot is employed to show the density of feature informativeness. Specifically, we run experiments 5 times and visualize all the feature informativeness from different modalities on the test datasets. From the experimental results in Fig. 5a, it is observed that the informativeness of most of the features is relatively low (e.g., close to 0), and



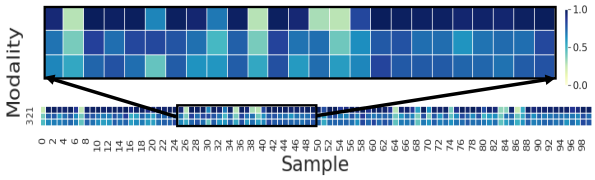
(a) Density of feature informativeness.

(b) Density of modality informativeness.

Figure 5. Density of the obtained feature and modality informativeness on BRCA dataset.



(a) Heatmap of feature informativeness.



(b) Heatmap of modality informativeness.

Figure 6. Heatmap of the obtained feature and modality informativeness on BRCA dataset.

only a small part of features are of greater informativeness (e.g., larger than 0.5). The reason for this is that the employed ℓ_1 loss could enforce the model to retain the most important features and eliminate the influence of unimportant features.

Density of modality informativeness. We further plot the density of the obtained modality informativeness with KDE to investigate the impact of the informativeness of modalities. For randomness issue, we run each experiment 5 times and show the obtained modality informativeness of different samples on the test dataset. The experimental results are shown in Fig. 5b. These different modalities of each sample have different informativeness. For example, mRNA expression of most samples are of high informativeness in decision-making, but there are also some samples whose mRNA expression modality informativeness are relatively low, which qualitatively illustrates the necessity of dynamically modeling the informativeness of modality in our method.

Heatmap of feature and modality informativeness. We further visualize the obtained feature and modality informativeness on BRCA dataset with heatmaps in Fig. 6a and Fig. 6b, respectively. We can observe that the proposed methods could perceive the dynamics of feature and modality for different samples. Specifically, in Fig. 6a, we have the following observations: (i) part of features are consistently uninformative on the different samples (close to 0); (ii) few features are important on all samples; (iii) the obtained informativeness of most features is constantly changed over different samples. Meanwhile, in Fig. 6b, the informativeness of different modalities is also changed dynamically over different samples due to factors such as noise and missing data during data collection.

5. Conclusion

In this paper, we propose a novel method termed Multimodal Dynamics for trustworthy multimodal classification. It can dynamically utilize informative features and modalities for each sample. To assess the informativeness of each feature, a sparse gating is introduced. Meanwhile the true class probability is employed to capture the informativeness dynamic in modality level. Then a dynamical fusion strategy is induced, which could provide a transparent fusion based on the informativeness of each feature and modality. Extensive experiments are performed on four multimodal medical classification datasets, where our method achieves superior classification performance and enhances trustworthiness and explainability.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61976151, 61732011), the National Key Research and Development Program of China under Grant 2019YFB2101900, the Natural Science Foundation of Tianjin of China (19JCYBJC15200), and the Key-Area Research and Development Program of Guangdong Province (2021B0101420005). We thank Tongxin Wang, the author of the paper [62], for his generous help in providing datasets.

References

[1] David A Bennett, Julie A Schneider, Zoe Arvanitakis, and Robert S Wilson. Overview and findings from the religious orders study. *Current Alzheimer Research*, 9(6):628–645, 2012. 5

[2] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 2021. 2

[3] Shotaro Akaho. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*, 2006. 1

[4] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013. 1

[5] Javier Antorán, James Urquhart Allingham, and José Miguel Hernández-Lobato. Depth uncertainty in neural networks. In *Advances in Neural Information Processing Systems*, 2020. 2

[6] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. In *ICLR workshop*, 2017. 1, 2, 5

[7] Ricardo Argelaguet. *Statistical methods for the integrative analysis of single-cell multi-omics data*. PhD thesis, University of Cambridge, 2020. 1

[8] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 1, 2

[9] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015. 2

[10] Anja-Katrin Bosserhoff, Markus Moser, Rüdiger Hein, Michael Landthaler, and Reinhard Buettner. In situ expression patterns of melanoma-inhibiting activity (mia) in melanomas and breast cancers. *The Journal of pathology*, 187(4):446–454, 1999. 7

[11] Charles Corbière, Nicolas THOME, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, 2019. 2, 4

[12] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 5

[13] Wenzhu Dai, Jixiang He, Ling Zheng, Mingyu Bi, Fei Hu, Minju Chen, Heng Niu, Jingyu Yang, Ying Luo, Wenru Tang, et al. mir-148b-3p, mir-190b, and mir-429 regulate cell progression and act as potential biomarkers for breast cancer. *Journal of breast cancer*, 22(2):219–236, 2019. 7

[14] Philip L De Jager, Yiyi Ma, Cristin McCabe, Jishu Xu, Badri N Vardarajan, Daniel Felsky, Hans-Ulrich Klein, Charles C White, Mette A Peters, Ben Lodgson, et al. A multi-omic atlas of the human frontal cortex for aging and alzheimer's disease research. *Scientific data*, 5(1):1–13, 2018. 5

[15] Vu C. Dinh and Lam Si Tung Ho. Consistent feature selection for analytic deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2420–2431, 2020. 3

[16] AROA Duro-Castano, Consuelo Borrás, Vicente Herranz-Pérez, M Carmen Blanco-Gandía, Inmaculada Conejos-Sánchez, Ana Armiñán, Cristina Mas-Bargues, Marta Inglés, Josse Miñarro, Marta Rodríguez-Arias, et al. Targeting alzheimer's disease with multimodal polypeptide-based nanoconjugates. *Science Advances*, 7(13):eabf9180, 2021. 1

[17] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019. 3

[18] Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3):238–247, 1989. 5

[19] Yarin Gal. Uncertainty in deep learning. 2016. 2

[20] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016. 2

[21] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021. 2

[22] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. 2017. 2

[23] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2020. 2, 5

[24] Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew Mingbo Dai, and Dustin Tran. Training independent subnetworks for robust prediction. In *International Conference on Learning Representations*, 2021. 2

[25] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995. 5

[26] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4340–4354, 2020. 1, 2, 5

[27] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1992. 1

[28] Han Hou, Qihao Zheng, Yuchen Zhao, Alexandre Pouget, and Yong Gu. Neural correlates of optimal multisensory decision making under time-varying reliabilities with an invariant linear probabilistic population code. *Neuron*, 104(5):1010–1021, 2019. 4

[29] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 1439–1449, October 2021. 1, 2

[30] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3507–3511, 2020. 2

[31] Peng Huang, Deng-jie Ouyang, Shi Chang, Mo-yun Li, Lun Li, Qian-ying Li, Rong Zeng, Qiong-yan Zou, Juan Su, Piao Zhao, et al. Chemotherapy-driven increases in the cdkn1a/ptn/ptprz1 axis promote chemoresistance by activating the nf-κb pathway in breast cancer cells. *Cell Communication and Signaling*, 16(1):1–12, 2018. 7

[32] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 5

[33] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019. 2

[34] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. *arXiv preprint arXiv:1802.02892*, 2018. 1

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2015. 5

[36] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017. 2

[37] Changhee Lee and Mihaela van der Schaar. A variational information bottleneck approach to multi-omics data integration. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 1513–1521, 2021. 2

[38] Hang Li, Fan Yang, Xiaohan Xing, Yu Zhao, Jun Zhang, Yueping Liu, Mengxue Han, Junzhou Huang, Liansheng Wang, and Jianhua Yao. Multi-modal multi-instance learning using weakly correlated histopathological images and tabular clinical information. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 529–539. Springer, 2021. 1, 2

[39] Huan Ma, Zongbo Han, Changqing Zhang, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. volume 34, 2021. 2

[40] Kleita Michaelidou, Alexandros Ardavanis, and Andreas Scorilas. Clinical relevance of the deregulated kallikrein-related peptidase 8 mrna expression in breast cancer: a novel independent indicator of disease-free survival. *Breast cancer research and treatment*, 152(2):323–336, 2015. 7

[41] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *International Conference on Machine Learning*, 2020. 4

[42] Laoighse Mulrane, Stephen F Madden, Donal J Brennan, Gabriela Gremel, Sharon F McGee, Sara McNally, Finian Martin, John P Crown, Karin Jirström, Desmond G Higgins, et al. mir-187 is an independent prognostic factor in breast cancer and confers increased invasive potential in vitro. *Clinical cancer research*, 18(24):6702–6713, 2012. 7

[43] Pradeep Natarajan, Shuang Wu, Shiv Vitaladevuni, Xiaodan Zhuang, Stavros Tsakalidis, Unsang Park, Rohit Prasad, and Premkumar Natarajan. Multimodal feature fusion for robust event detection in web videos. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1298–1305. IEEE, 2012. 2

[44] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012. 2

[45] Rameswar Panda, Chun-Fu (Richard) Chen, Quanfu Fan, Ximeng Sun, Kate Saenko, Aude Oliva, and Rogerio Feris. Adamml: Adaptive multi-modal learning for efficient video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7576–7585, October 2021. 1

[46] Rameswar Panda, Chun-Fu Richard Chen, Quanfu Fan, Ximeng Sun, Kate Saenko, Aude Oliva, and Rogerio Feris. Adamml: Adaptive multi-modal learning for efficient video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7576–7585, 2021. 2

[47] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 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*, pages 2539–2544, 2015. 2

[48] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017. 2

[49] Reuben Rideaux, Katherine R Storrs, Guido Maiello, and Andrew E Welchman. How multisensory neurons solve causal inference. *Proceedings of the National Academy of Sciences*, 118(32), 2021. 4

[50] Manuel Scimeca, Chiara Antonacci, Nicola Toschi, Elena Giannini, Rita Bonfiglio, Claudio Oreste Buonomo, Chiara Adriana Pistolese, Umberto Tarantino, and Elena Bonanno. Breast osteoblast-like cells: a reliable early marker for bone metastases from breast cancer. *Clinical breast cancer*, 18(4):e659–e669, 2018. 7

[51] Seonguk Seo, Paul Hongsuck Seo, and Bohyung Han. Learning for single-shot confidence calibration in deep neural networks through stochastic inferences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9030–9038, 2019. 2

[52] Paulina Siejka-Zielińska, Jingfei Cheng, Felix Jackson, Yibin Liu, Zahir Soonawalla, Srikanth Reddy, Michael Silva, Luminita Puta, Misti Vanette McCain, Emma L Culver, et al. Cell-free dna taps provides multimodal information for early cancer detection. *Science advances*, 7(36):eabh0534, 2021. 1

[53] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014. 2

[54] Amrit Singh, Casey P Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J Tebbutt, and Kim-Anh Lê Cao. Diablo: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17):3055–3062, 2019. 5

[55] Mahesh Subedar, Ranganath Krishnan, Paulo Lopez Meyer, Omesh Tickoo, and Jonathan Huang. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6301–6310, 2019. 2

[56] Junjiao Tian, Wesley Cheung, Nathaniel Glaser, Yen-Cheng Liu, and Zsolt Kira. Uno: Uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation. In *2020 IEEE International Conference on Robotics and Automation*, pages 5716–5723. IEEE, 2020. 2

[57] Ashwini Tonge and Cornelia Caragea. Dynamic deep multimodal fusion for image privacy prediction. In *The World Wide Web Conference*, pages 1829–1840, 2019. 2

[58] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019. 2

[59] Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, 2020. 4

[60] Mark A Van De Wiel, Tonje G Lien, Wina Verlaat, Wessel N van Wieringen, and Saskia M Wilting. Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in Medicine*, 35(3):368–381, 2016. 5

[61] Hao Wang, Yan Yang, and Bing Liu. GMC: Graph-based multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 32(6):1116–1129, 2020. 1

[62] Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, 12(1):1–13, 2021. 1, 2, 5, 8

[63] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092, 2015. 1

[64] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 1

[65] Yang Wang. Survey on deep multi-modal data analytics: collaboration, rivalry, and fusion. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1s):1–25, 2021. 1, 2

[66] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[67] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning*, pages 681–688, 2011. 2

[68] Yukako Yagi. Color standardization and optimization in whole slide imaging. In *Diagnostic pathology*, volume 6, pages 1–12. Springer, 2011. 1

[69] Akimitsu Yamada, Takashi Ishikawa, Ikuko Ota, Mariko Kimura, Daisuke Shimizu, Mikiko Tanabe, Takashi Chishima, Takeshi Sasaki, Yasushi Ichikawa, Satoshi Morita, et al. High expression of atp-binding cassette transporter abcc11 in breast tumors is associated with aggressive subtypes and low disease-free survival. *Breast cancer research and treatment*, 137(3):773–782, 2013. 7

[70] UshaRani Yelipe, Sammulal Porika, and Madhu Golla. An efficient approach for imputation and classification of medical data values using class-based clustering of medical records. *Computers & Electrical Engineering*, 66:487–504, 2018. 1

[71] Changqing Zhang, Zongbo Han, Huazhu Fu, Joey Tianyi Zhou, Qinghua Hu, et al. Cpm-nets: Cross partial multi-view networks. In *Advances in Neural Information Processing Systems*, pages 559–569, 2019. 1

[72] Heng Zhang, Vishal M Patel, and Rama Chellappa. Hierarchical multimodal metric learning for multimodal classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3057–3065, 2017. 1

[73] Jian Zhang, Jianli Luo, Huali Jiang, Tao Xie, Jieling Zheng, Yunhong Tian, Rong Li, Baiyao Wang, Jie Lin, Anan Xu, et al. The tumor suppressor role of zinc finger protein 671 (znf671) in multiple tumors based on cancer single-cell sequencing. *Frontiers in oncology*, 9:1214, 2019. 7

[74] Pengfei Zhu, Xinjie Yao, Yu Wang, Meng Cao, Binyuan Hui, Shuai Zhao, and Qinghua Hu. Latent heterogeneous graph network for incomplete multi-view learning. *IEEE Transactions on Multimedia*, 2022. 1