CLIMD: A Curriculum Learning Framework for Imbalanced Multimodal Diagnosis

Kai Han, Chongwen Lyu, Lele Ma, Chengxuan Qian,
Siqi Ma, Zheng Pang, Jun Chen, and Zhe Liu ✉

School of Computer Science and Telecommunication Engineering,
Jiangsu University, China
1000004088@ujs.edu.cn

Abstract. Clinicians usually combine information from multiple sources to achieve the most accurate diagnosis, and this has sparked increasing interest in leveraging multimodal deep learning for diagnosis. However, in real clinical scenarios, due to differences in incidence rates, multimodal medical data commonly face the issue of class imbalance, which makes it difficult to adequately learn the features of minority classes. Most existing methods tackle this issue with resampling or loss reweighting, but they are prone to overfitting or underfitting and fail to capture cross-modal interactions. Therefore, we propose a **C**urriculum **L**earning (CL) framework for **I**mbalanced **M**ultimodal **D**iagnosis (CLIMD). Specifically, we first design multimodal curriculum measurer that combines two indicators, intra-modal confidence and inter-modal complementarity, to enable the model to focus on key samples and gradually adapt to complex category distributions. Additionally, a class distribution-guided training scheduler is introduced, which enables the model to progressively adapt to the imbalanced class distribution during training. Extensive experiments on multiple multimodal medical datasets demonstrate that the proposed method outperforms state-of-the-art approaches across various metrics and excels in handling imbalanced multimodal medical data. Furthermore, as a plug-and-play CL framework, CLIMD can be easily integrated into other models, offering a promising path for improving multimodal disease diagnosis accuracy. Code is publicly available at https://github.com/KHan-UJS/CLIMD.

Keywords: Multimodal deep learning · Class imbalance · Curriculum learning · Computer-aided diagnosis.

1 Introduction

Thanks to the rapid development of deep learning, Computer-Aided Diagnosis (CAD) has experienced a new era of advancement. Traditional deep diagnosis researches mostly focus on a single modality [6, 9, 12, 19], however, diagnosing diseases based on only one modality is challenging and risky. Therefore, in recent years, the application of multimodal deep learning (MMDL) in disease diagnosis,

which makes comprehensive judgments by combining information from multiple modalities, has garnered widespread attention [10, 14, 21].

In real medical cases, common diseases usually constitute a significant proportion, while samples of some rare malignant diseases are relatively scarce, leading to the issue of class imbalance. This issue makes it challenging for the model to learn the minority classes, which can lead to biased predictions. There is relatively limited research addressing the class imbalance issue in multimodal diagnosis (MD). Some existing methods attempt to apply common resampling [4], which carries the risk of overfitting or underfitting, as well as loss reweighting [13, 15, 23], which may increase the instability of model training. Additionally, these methods overlook the impact of interactions between modalities in MMDL.

To this end, we propose a MD framework based on curriculum learning (CL) [1, 18] named CLIMD. CL enables the model to transition gradually from simple to complex during the training process [3]. Inspired by the concept of CL, we attempted to quantify the degree of class imbalance in the dataset and then train the model using a progressive sampling strategy. Specifically, we first propose a curriculum measurer combining intra-modal confidence and inter-modal complementarity, to evaluate the training difficulty of each multimodal sample. Based on the difficulty, a class distribution-guided curriculum scheduling algorithm is designed to quantify the degree of class imbalance and construct training subsets with gradually increasing class imbalance in each epoch, helping the model adapt to imbalanced data. CLMD does not involve any down-sampling, data generation, or data augmentation operations, thereby avoiding the issues of underfitting or overfitting arise from them. The main contributions of this paper can be summarized as follows:

- We propose a novel framework based on curriculum learning to mitigate the issue of class imbalance in multimodal diagnosis. To the best of our knowledge, this is the first CL strategy specifically designed for imbalanced multimodal diagnosis.
- We design a curriculum measurer and a class distribution-guided training scheduler to measure the training difficulty of multimodal samples and help the model gradually adapt to the imbalance during training.
- Extensive experiments conducted on multiple datasets demonstrate that the proposed method significantly outperforms state-of-the-art methods and can effectively address the issue of class imbalance.

2   Method

In this section, we introduce the proposed method in detail, including a Multimodal Curriculum Measurer (MCM) and a Training Sample Scheduler (TSS). The overview of CLIMD is shown in Fig. 1. It is noteworthy that, to evaluate the upper bound of the proposed CL without interference from other factors, the multimodal diagnosis (MD) network we employ (i.e., **Baseline**) is a basic and classical early fusion architecture with Cross-Entropy (CE) as the loss function.
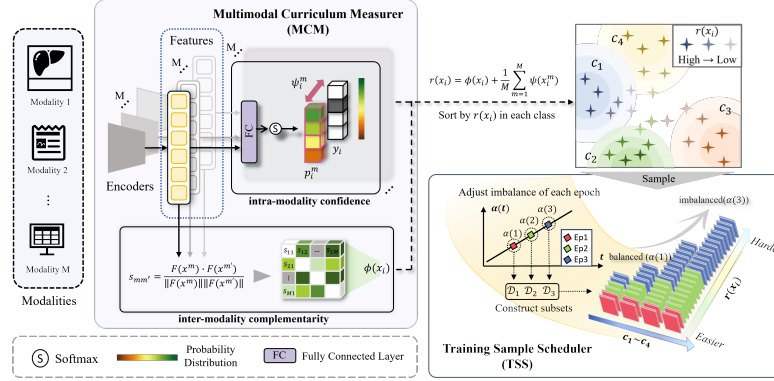
Fig. 1. Overview of CLIMD, which includes Multimodal Curriculum Measurer (MCM) and Training Sample Scheduler (TSS). (i) MCM evaluates the training difficulty of each sample from intra-modal confidence $\psi_i^{(m)}$ and inter-modal complementarity $\phi(x_i)^{(m)}$, which are combined to calculate the overall training difficulty $r(x_i)$ for each sample. (ii) TSS samples from each class in order to construct the training subset $\mathcal{D}_t$ for the t-th epoch and gradually increases the class imbalance degree in the training subset by adjusting $\alpha(t)$.

In fact, CLIMD is plug-and-play and can be applied to other more complex MD models.

2.1   Definition

Given a dataset $\mathcal{D}$ with M modalities, which includes N multimodal data observations with corresponding labels, i.e., $\mathcal{D} = \{\{x_i^{(m)}\}_{m=1}^M, y_i\}_{i=1}^N$. The main goal of the MD task is to learn a mapping from the multimodal data space $\boldsymbol{X} = \{x_i^{(m)} \in \mathbb{R}^{d_m}\}_{m=1}^M$ to the label space $\boldsymbol{Y} = \{y_i \in \mathbb{R}^C\}$, where d_m and C are the dimensionality of M modalities and class number, respectively.

2.2   Multimodal Curriculum Measurer

In imbalanced multimodal learning, random sampling may lead to model bias toward majority classes or insufficient consideration of modality interactions. To sample in a specific order, we propose the multimodal curriculum measurer as a structured evaluation criterion to regulate the training process. The Multimodal Curriculum Measurer (MCM) comprises intra-modal confidence, which prioritizes stable samples to enhance convergence robustness, and inter-modal complementarity, which facilitates the learning of complementary information across modalities to improve fusion effectiveness.

Intra-modal Confidence The prediction confidence of a sample in its m-th modality quantifies the model's certainty regarding the true class. To compute

this, the class probability output $p_{i,c}^m$ and the true class label $y_{i,c}$ are utilized to derive the confidence score $\psi(x_i^m)$, as defined in Eq. 1.

$$\psi(x_i^m) = \sigma \left[\frac{1}{C} \sum_{c=1}^{C} y_{i,c} \log p_{i,c}^m \right] \tag{1}$$

where σ denotes the sigmoid function.

Inter-modal Complementarity Inter-modal complementarity quantifies the complementarity score by computing modality feature similarity. Specifically, we extract modality-specific representations $F(x^m)$ through encoders, then compute pairwise modality feature similarity $s_{mm'}$ to obtain the similarity matrix $\mathcal{S}$. $s_{mm'} \in \mathcal{S}$, where $m \in \{1, \cdots, M\}$, $m' \in \{1, \cdots, M\}$. The calculation process of feature similarity $s_{mm'}$ is as follows:

$$s_{mm'} = \frac{F(x^m) \cdot F(x^{m'})}{\|F(x^m)\| \|F(x^{m'})\|} \tag{2}$$

Based on the similarity matrix $\mathcal{S}$, modality complementarity is computed as shown in the following equation.

$$\phi(x_i) = 1 - \frac{1}{M(M-1)} \sum_{m \neq m'} s_{mm'} \tag{3}$$

Finally, the intra-modal confidence and inter-modal complementarity scores are integrated, as shown in Eq. 4.

$$r(x_i) = \phi(x_i) + \frac{1}{M} \sum_{m=1}^{M} \psi(x_i^m) \tag{4}$$

2.3   Class Distribution-guided Training Scheduler

To help the model gradually adapt to imbalance, we propose a Training Sample Scheduler (TSS) guided by class distribution. TSS first quantifies the degree of class imbalance in the dataset. Then, the model starts learning from a subset with relatively balanced distribution, and progressively increases the degree of class imbalance until all imbalanced data is involved in the training process.

Fit the Class Distribution We assume that there are C classes in the dataset $\mathcal{D}$, i.e., $\mathcal{D} = \{\mathcal{D}_{sub}^{(c)}\}_{c=1}^{C}$. A discrete class distribution $N = \{n^{(c)}\}_{c=1}^{C}$ can be obtained, where $n^{(c)}$ represents the number of samples in the c-th class. Next, we fit N to an improved power-law distribution, whose general probability density function (PDF) is as follows:

$$P(N) = (\gamma\alpha - 1)n_{min}^{\gamma\alpha-1} N^{-\gamma\alpha} \tag{5}$$
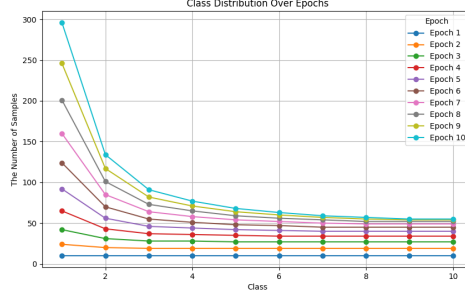
Fig. 2. An example of class distribution over epochs, where the sample number $N = 1000$, the total epochs $T = 10$, the number of classes $C = 10$, the hyperparameter of the real class distribution (in Epoch 10) function $\alpha(T) = 5$.

where n_{min} represents the smallest value in N, $\gamma > 0$ is a hyperparameter that makes the distribution function smoother, and $\alpha > 1/\gamma$ is the parameter to be estimated. The optimal solution for α is obtained through Maximum Likelihood Estimation (MLE) as:

$$\hat{\alpha} = \frac{1}{\gamma} \left(1 + \frac{C}{\sum_{c=1}^{C} \ln n^{(c)} - C \ln n_{min}} \right) \tag{6}$$

Notably, $\hat{\alpha}$ here can be regarded as the upper bound of the class imbalance degree, and it is written as $\alpha(T)$ in the following content.

Construct the Training Subsets According to training difficulty $r(x_i)$ obtained in Section 2.2, all samples within their respective classes are sorted from easy to hard, waiting to be sampled in order later. We assume that the total number of training epochs is T, and the training subset for the t-th epoch is $\mathcal{D}_t$, where $t \in [0, T]$. The number of samples in each training subset is $S_t = \left\{ s_t^{(c)} \right\}_{c=1}^{C} = \frac{t}{T} \cdot N$, where $s_t^{(c)}$ represents the number of samples in the c-th class. In the first training subset $\mathcal{D}_1$, the classes follow a uniform distribution, meaning that the number of samples for each class is S_1/C. Next, we control the degree of class imbalance in each epoch by adjusting $\alpha(t) = 1 + (\alpha(T) - 1) \cdot \frac{t-1}{T-1}$, and $1 \le \alpha(t) \le \alpha(T)$. Therefore, the class distribution function of the training subset in the t-th epoch is a mixture of uniform distribution and power-law distribution, as follows:

$$q_t(c) = \left(1 - \frac{t-1}{T-1} \right) \cdot \frac{1}{C} + \left(\frac{t-1}{T-1} \right) \cdot \frac{c^{-\gamma\alpha(t)}}{\sum_{i=1}^{C} i^{-\gamma\alpha(t)}} \tag{7}$$

where $q_t(c)$ represents the sampling probability of c-th class in the t-th epoch, and the number of samples drawn from the c-th class in order is $s_t^{(c)} = q_t(c) \cdot S_t$.

To provide an intuitive illustration, we have included an example in Fig. 2, where $N = 1000, T = 10, C = 10, \alpha(T) = 5$. It shows that the proposed scheduling strategy allows the training subsets to gradually transition from a uniform distribution to a long-tail distribution (which approximates real imbalanced class distribution), enabling the model to progressively adapt to the class imbalance in the training samples.

2.4   Experiment

Datasets and Evaluation Metrics Multimodal Liver Lesion Classification (MLLC) is a private multimodal liver lesion diagnosis dataset containing 320 patients, involving three classes: normal, metastatic liver lesions, and hepatocellular carcinoma. It consists of CT arterial phase images, medical history texts, and laboratory test indicators. Breast Invasive Carcinoma (BRCA) is a publicly available dataset in The Cancer Genome Atlas (TGCA) [1] program. It is used for breast invasive carcinoma subtype classification, containing 3 modalities: mRNA, DNA methylation, and miRNA.

Three metrics are used to quantify the experimental results: accuracy (ACC), weighted F1 score, and macro F1 score.

Implementation Details For the MLLC dataset, a pretrained ResNet18 [8] is used to extract features from CT. In addition, we use MedBERT [20] to encode medical texts and a two-layer fully connected network (FCN) to encode the tabular data. For the BRCA dataset, we preprocess it referring to [22] [2]. The learning rate is set to 5e-5, the training epochs T is 100, the hyperparameter γ is empirically set to 0.3. All training processes are performed on an NVIDIA GeForce RTX A6000 GPU.

Table 1. The comparisons with SOTA MD methods on MLLC and BRCA.

Method	MLLC			BRCA		
	ACC	Weighted F1	Macro F1	ACC	Weighted F1	Macro F1
MORONET [22] (NC' 2021)	72.2 ± 4.6	67.2 ± 4.9	68.3 ± 6.8	82.9 ± 1.8	82.5 ± 1.6	77.4 ± 1.7
Multi-D [7] (CVPR' 2022)	74.4 ± 5.0	68.1 ± 4.6	70.3 ± 4.2	87.7 ± 0.3	88.0 ± 0.5	84.5 ± 0.5
MLCLNet [25] (AAAI' 2023)	76.4 ± 3.5	74.5 ± 3.2	72.6 ± 3.7	86.4 ± 1.6	87.8 ± 1.7	82.6 ± 1.8
DPNET [27] (ACM MM' 2023)	76.7 ± 4.3	72.7 ± 3.7	73.8 ± 4.0	87.8 ± 1.0	88.4 ± 1.2	85.2 ± 1.2
DMIB [5] (WACV' 2024)	77.6 ± 3.6	71.3 ± 4.3	74.9 ± 3.2	86.0 ± 0.7	86.0 ± 0.9	81.6 ± 0.9
PCAG [24] (NN' 2024)	78.1 ± 3.1	73.0 ± 3.1	75.6 ± 3.1	85.2 ± 1.7	85.5 ± 1.9	81.4 ± 2.6
GCFANet [26] (IF' 2024)	76.5 ± 3.9	73.5 ± 2.4	77.8 ± 2.7	88.6 ± 1.5	88.9 ± 1.6	85.3 ± 1.6
CLIMD	$\mathbf{78.5 \pm 2.3}$	$\mathbf{75.4 \pm 2.5}$	$\mathbf{78.5 \pm 2.3}$	$\mathbf{89.2 \pm 1.4}$	$\mathbf{89.4 \pm 1.2}$	$\mathbf{86.5 \pm 1.8}$

Comparing with SOTA Methods We compare CLIMD with state-of-the-art MD methods [5, 7, 22, 25–27], and the results are shown in Table 1. CLIMD achieves the best performance on both datasets, demonstrating the effectiveness of the proposed curriculum metric in evaluating the degree of imbalance among training samples.

Table 2. The comparisons with other methods specifically designed for class imbalance on MLLC and BRCA.

Method	MLLC			BRCA		
	ACC	Weighted F1	Macro F1	ACC	Weighted F1	Macro F1
Random Undersampling	73.2 ± 5.1	69.3 ± 5.2	71.6 ± 6.3	83.6 ± 1.4	84.2 ± 0.8	81.1 ± 1.5
Oversampling (SMOTE [2])	70.2 ± 6.8	70.3 ± 5.0	68.8 ± 7.4	79.7 ± 1.3	80.4 ± 1.0	76.6 ± 1.6
Oversampling (VAE)	74.4 ± 4.7	72.3 ± 4.3	71.6 ± 5.6	82.4 ± 1.6	83.4 ± 2.1	79.2 ± 1.8
WCE Loss	75.5 ± 4.7	74.0 ± 3.1	72.5 ± 4.3	84.8 ± 2.2	85.0 ± 2.2	81.7 ± 2.5
Focal Loss [11]	75.8 ± 2.8	74.7 ± 3.7	75.8 ± 3.8	85.1 ± 1.2	85.6 ± 0.8	82.6 ± 1.1
CLIMD	**78.5 ± 2.3**	**75.4 ± 2.5**	**78.5 ± 2.3**	**89.2 ± 1.4**	**89.4 ± 1.2**	**86.5 ± 1.8**

Evaluation of Addressing Class Imbalance To evaluate the performance of CLIMD in addressing the issue of class imbalance, we compare it with other methods specifically designed for class imbalance, and the results are presented in Table 2. The explanation for each method is as follows: (1) **Random Undersampling:** Reducing the number of majority classes by random downsampling; (2) **Oversampling (SMOTE):** Using SMOTE [2] algorithm to oversample the minority classes; (3) **Oversampling (VAE):** Generating new samples by VAE to oversample the minority classes; (4) **Weighted Cross-Entropy (WCE) Loss:** Assigning different weights to different class samples using the WCE loss; (5) **Focal Loss:** Dynamically adjusting the contribution of each sample to the loss using focal loss [11]. The results demonstrate that CLIMD is more efficient in addressing the issue of class imbalance.

Improvements to Different Models Our proposed CL framework is plug-and-play, which implies its potential applicability to various other MD models to further improve their performance. To substantiate this claim, we conducted experiments based on several state-of-the-art models [5, 7], and the results are shown in Table 3. It can be seen that after applying CLIMD, the performance of various MD models has improved to varying degrees. Moreover, the improvement is more pronounced for simpler models.

Ablation Study A series of ablation studies are conducted to validate the effectiveness of the proposed MCM and TSS, the results are shown in Table 4. Methods (1) and (2) only use intra-modal or inter-modal measure, respectively.

Table 3. The improvement of applying CLIMD on BRCA dataset.

Method	ACC	Weighted F1	Macro F1
Baseline	83.6 ± 1.6	84.4 ± 1.5	81.3 ± 2.1
Baseline (*w/* CLIMD)	89.2 ± 1.4	89.4 ± 1.2	86.5 ± 1.8
Multi-D	87.7 ± 0.3	88.0 ± 0.5	84.5 ± 0.5
Multi-D (*w/* CLIMD)	89.6 ± 0.6	90.2 ± 1.0	85.4 ± 0.8
DMIB	86.0 ± 0.7	86.0 ± 0.8	81.6 ± 0.9
DMIB (*w/* CLIMD)	88.4 ± 0.6	88.9 ± 1.1	83.8 ± 1.4

Table 4. The ablation study of MCM and TSS on BRCA.

Method	Curriculum Measurer		Training Scheduler	ACC	Weighted F1	Macro F1
	Intra-modal	Inter-modal				
Baseline	-	-	-	83.6 ± 1.6	84.4 ± 1.5	81.3 ± 2.1
(1)	✓	-	ours	86.6 ± 1.1	86.9 ± 0.7	83.4 ± 1.4
(2)	-	✓	ours	87.5 ± 1.2	88.0 ± 0.9	84.3 ± 1.5
(3)	✓	✓	Baby Step	85.1 ± 2.4	86.1 ± 2.1	80.2 ± 2.7
(4)	✓	✓	One-Pass	84.9 ± 1.6	85.6 ± 1.2	79.2 ± 1.9
(5)	✓	✓	Linear	86.6 ± 2.1	87.3 ± 1.6	84.0 ± 2.4
(6)	✓	✓	Root	88.3 ± 1.6	88.8 ± 1.1	85.4 ± 1.6
(7)	✓	✓	Geom	87.7 ± 1.9	88.3 ± 1.5	84.8 ± 2.2
CLIMD	✓	✓	ours	$\mathbf{89.2 \pm 1.4}$	$\mathbf{89.4 \pm 1.2}$	$\mathbf{86.5 \pm 1.8}$

Methods (3)-(7) employ Baby Step [1], One-Pass [1], Linear, Root [17], and Geometric Progression [16] as training schedulers, respectively. The ablation experiments reveal that the optimal results are achieved when both the proposed curriculum measures and training scheduler are utilized.

3   Conclusion

In this paper, a novel CL framework is proposed to address the issue of class imbalance in multimodal diagnosis. We first leverage intra-modal confidence and inter-modal complementarity to measure training difficulty of each multimodal sample. Then, we design a novel scheduler based on class distribution, which aims to lead the model to gradually adapt to the imbalanced class distribution. Extensive experimental results on multiple datasets demonstrate the efficiency of the proposed method. In future work, we will evaluate the proposed method across diverse medical scenarios using additional datasets.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this paper.

References

1. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009)
2. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
3. Elman, J.L.: Learning and development in neural networks: The importance of starting small. Cognition **48**(1), 71–99 (1993)
4. Elola, A., Aramendi, E., Irusta, U., Berve, P.O., Wik, L.: Multimodal algorithms for the classification of circulation states during out-of-hospital cardiac arrest. IEEE Transactions on Biomedical Engineering **68**(6), 1913–1922 (2020)
5. Fang, Y., Wu, S., Zhang, S., Huang, C., Zeng, T., Xing, X., Walsh, S., Yang, G.: Dynamic multimodal information bottleneck for multimodality classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7696–7706 (2024)
6. Ge, Y., Guo, Y., Das, S., Al-Garadi, M.A., Sarker, A.: Few-shot learning for medical text: A review of advances, trends, and opportunities. Journal of Biomedical Informatics p. 104458 (2023)
7. Han, Z., Yang, F., Huang, J., Zhang, C., Yao, J.: Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20707–20717 (2022)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Hu, R., Song, Y., Liu, Y., Zhu, Y., Feng, N., Qiu, C., Han, K., Teng, Q., Haq, I.U., Liu, Z.: Imbalance multiclass problem: a robust feature enhancement-based framework for liver lesion classification. Multimedia Systems **30**(2), 104 (2024)
10. Huang, S.C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P.: Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ digital medicine **3**(1), 136 (2020)
11. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
12. Liu, Z., Han, K., Xue, K., Song, Y., Liu, L., Tang, Y., Zhu, Y.: Improving ct-image universal lesion detection with comprehensive data and feature enhancements. Multimedia Systems **28**(5), 1741–1752 (2022)
13. Mallya, M., Hamarneh, G.: Deep multimodal guidance for medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 298–308. Springer (2022)
14. Mohsen, F., Ali, H., El Hajj, N., Shah, Z.: Artificial intelligence-based methods for fusion of electronic health records and imaging data. Scientific Reports **12**(1), 17981 (2022)
15. Morais, M., Calisto, F.M., Santiago, C., Aleluia, C., Nascimento, J.C.: Classification of breast cancer in mri with multimodal fusion. In: 2023 IEEE 20th international symposium on biomedical imaging (ISBI). pp. 1–4. IEEE (2023)
16. Penha, G., Hauff, C.: Curriculum learning strategies for ir: An empirical study on conversation response ranking. In: Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42. pp. 699–713. Springer (2020)

17. Platanios, E.A., Stretcu, O., Neubig, G., Poczos, B., Mitchell, T.M.: Competence-based curriculum learning for neural machine translation. arXiv preprint arXiv:1903.09848 (2019)
18. Qian, C., Han, K., Wang, J., Yuan, Z., Lyu, C., Chen, J., Liu, Z.: Dyncim: Dynamic curriculum for imbalanced multimodal learning. arXiv preprint arXiv:2503.06456 (2025)
19. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine **4**(1), 86 (2021)
20. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine **4**(1), 86 (2021)
21. Sun, Z., Lin, M., Zhu, Q., Xie, Q., Wang, F., Lu, Z., Peng, Y.: A scoping review on multimodal deep learning in biomedical images and texts. Journal of Biomedical Informatics p. 104482 (2023)
22. Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., Huang, K.: Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nature communications **12**(1), 3445 (2021)
23. Yue, G., Wei, P., Liu, Y., Luo, Y., Du, J., Wang, T.: Automated endoscopic image classification via deep neural network with class imbalance loss. IEEE Transactions on Instrumentation and Measurement **72**, 1–11 (2023)
24. Zhang, D., Nayak, R., Bashar, M.A.: Pre-gating and contextual attention gate–a new fusion method for multi-modal data tasks. Neural Networks p. 106553 (2024)
25. Zheng, X., Tang, C., Wan, Z., Hu, C., Zhang, W.: Multi-level confidence learning for trustworthy multimodal classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 11381–11389 (2023)
26. Zheng, X., Wang, M., Huang, K., Zhu, E.: Global and cross-modal feature aggregation for multi-omics data classification and application on drug response prediction. Information Fusion **102**, 102077 (2024)
27. Zou, X., Tang, C., Zheng, X., Li, Z., He, X., An, S., Liu, X.: Dpnet: Dynamic poly-attention network for trustworthy multi-modal classification. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 3550–3559 (2023)