Learning Contrastive Multimodal Fusion with Improved Modality Dropout for Disease Detection and Prediction

Yi Gu[1,2], Kuniaki Saito[1], and Jiaxin Ma[1(✉)]

[1] OMRON SINIC X Corporation
jiaxin.ma@sinicx.com
[2] Nara Institute of Science and Technology

Abstract. As medical diagnoses increasingly leverage multimodal data, machine learning models are expected to effectively fuse heterogeneous information while remaining robust to missing modalities. In this work, we propose a novel multimodal learning framework that integrates enhanced modalities dropout and contrastive learning to address real-world limitations such as modality imbalance and missingness. Our approach introduces learnable modality tokens for improving missingness-aware fusion of modalities and augments conventional unimodal contrastive objectives with fused multimodal representations. We validate our framework on large-scale clinical datasets for disease detection and prediction tasks, encompassing both visual and tabular modalities. Experimental results demonstrate that our method achieves state-of-the-art performance, particularly in challenging and practical scenarios where only a single modality is available. Furthermore, we show its adaptability through successful integration with a recent CT foundation model. Our findings highlight the effectiveness, efficiency, and generalizability of our multimodal learning approach, offering a scalable, low-cost solution with significant potential for more complicated clinical applications that allow missing modality input. The code is available at https://github.com/omron-sinicx/medical-modality-dropout.

Keywords: Neural network fusion · Contrastive learning · Chest computed tomography (CT) · Lung diseases.

1 Introduction

Advancements in medical diagnostic have led to increasingly diverse clinical data, comprising multimodal sources such as medical images (e.g., computed tomography [CT] and magnetic resonance imaging) and tabular data (e.g., electronic health records [EHRs] and radiology reports). This diversity has driven research in multimodal learning to enhance predictive modeling in health care [1,7,24,29,2,4,12,30,18,26,31,37,6,3,36,11,32]. Deep learning models trained on multimodal data have demonstrated improved performance in disease detection

and prediction, including lung diseases analysis [15,8,38]. Despite these advancements, effectively leveraging multimodal information remains challenging due to real-world constraints such as modality imbalance and missingness.

Previous works have attempted to mitigate these challenges through modality dropout, wich enhances model robustness by simulating missing modalities during training [10,23,16,25]. However, the traditional modality dropout uses fixed placeholders, limiting its ability to improve missingness awareness. Recent methods have explored learnable missingness instructions [1,32,17]; however, their integration into modality dropout remains underexplored. On the other hand, contrastive learning has emerged as a powerful technique for multimodal representation learning [7,29,36,4,12,31,28,14,35,34]. By encouraging models to associate information from heterogeneous sources that refer to the same underlying concept (e.g., the same patient or event), contrastive learning improved downstream task performance. However, most contrastive methods focus on unimodal representations, whereas the fused multimodal representations were not utilized.

In this research, we propose a novel multimodal learning framework designed to enhance both unimodal and multimodal performance for disease detection and prediction. We build a multimodal model consisting of unimodal encoders, a neural fusion module, and a task-specific head. We assume the unimodal encoders are pretrained and frozen during our training to demonstrate the improvement with a low cost of additional training. Unlike previous work that produced unimodal representations by unimodal encoders, our method leverages the multimodal model to generate unimodal representations with modality dropout. Additionally, we introduce learnable modality tokens in modality dropout to improve the model's awareness of missing modalities. Furthermore, we propose multimodal contrastive learning with fused multimodal representations for better representation binding. We validate our method using two large-scale public clinical datasets with three tasks of disease detection and prediction from CT and tabular data. We also employ a recent CT foundation model [33] as the encoder and show the improvement by our method. **The contribution** of this work is threefold: 1) We propose a novel multimodal learning framework with improved modality dropout and contrastive multimodal learning. 2) We demonstrate the effectiveness and efficacy of the proposed method using large-scale public clinical datasets for disease detection and prediction. 3) We show the efficient improvement on a recent CT foundation model at a low cost.

2 Method

The proposed multimodal learning framework is illustrated in Fig. 1. We aim to train a multimodal model $F_\theta(\cdot,\cdot)$ (parameterized by θ) that detects or predicts diseases from medical images and tabular data while being robust to missing modalities during inference. Our model comprises pretrained unimodal encoders for processing individual modalities, a lightweight fusion module of multilayer perception (MLP) to integrate information from different sources, and a task-

specific head, which is a classifier for the target training (Fig. 1 [a]) or a projector for pretraining (Fig. 1 [b]). The unimodal encoders remain frozen throughout all training to reduce additional training overhead.

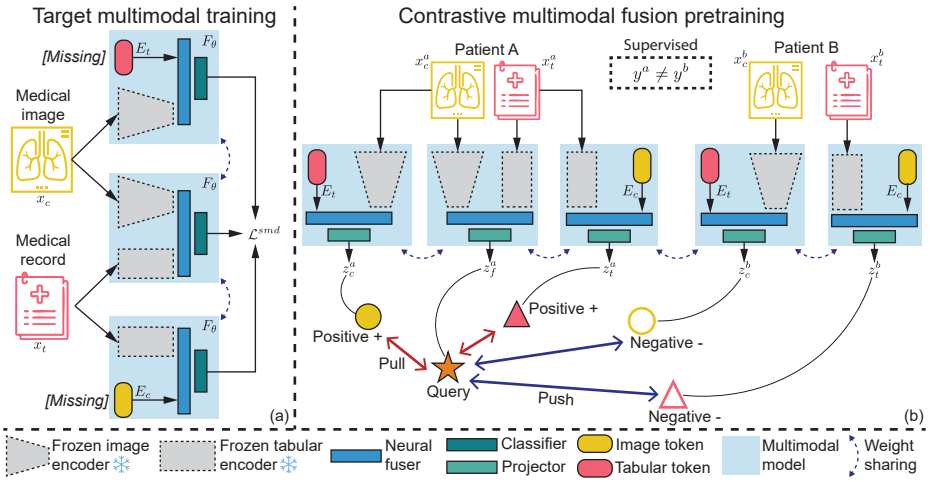

Fig. 1. Overview of the proposed method for training the multimodal model. (a) Target training using the simultaneous modality dropout, where the multimodal model is supervised using unimodal and multimodal inputs simultaneously. (b) Inter-modality contrastive learning using the multimodal and unimodal representations. The learnable modality tokens are introduced for improving missingness-aware neural fusion.

Simultaneous modality dropout. The modality dropout simulates that modality is missing for a patient during training. We consider the task of detecting or predicting disease from medical images and tabular data as illustrated in Fig. 1 (a). Given a patient sample, with image x_c^i and tabular x_t^i modalities, the multimodal model $F_\theta(\cdot,\cdot)$ predicts the probability $p(y^i \mid x_c^i, x_t^i, \theta) = F_\theta(x_c^i, x_t^i)$, where y^i is the associated label. The model is trained by maximizing the log-likelihood using $\mathcal{L}^{base} = -\log p(y^i \mid x_c^i, x_t^i, \theta)$. We adopt modality dropout [10,23,16,25] to enable robustness to missing modalities. Unimodal predictions, $p(y^i \mid x_c^i, \theta) = F_\theta(x_c^i, \mathbf{0}_t)$ and $p(y^i \mid x_t^i, \theta) = F_\theta(\mathbf{0}_c, x_t^i,)$, are obtained by replacing missing modalities with zero matrices, where the $\mathbf{0}_c$ and $\mathbf{0}_t$ are the zero metrics for the image and tabular modalities, respectively. Traditional modality dropout training $\mathcal{L}^{md} = -\log p(y^i \mid \{x_j^i\}_{j\in S}, \theta)$ employs a random sampling function $g(\cdot)$ to select a subset $S = \{g(M) \in \mathcal{P}(M) \setminus \varnothing\}$ of available modalities at each iteration, where $\mathcal{P}(M)$ represents the power of all modalities $M = \{c, t\}$. The sampling function $g(\cdot)$ was introduced to avoid large computational costs in a single iteration as the number of modality combinations $|\mathcal{P}(M)| = 2^{|M|}$ scales exponentially [23]. Instead, we propose simultaneous modality dropout, where

all modality combinations are explicitly supervised without sampling, leveraging the small number of modalities and the lightweight nature of our fusion module. Our loss function is defined as Eq. 1, where the λ is a hyperparameter that balances term weights. This way, our approach ensures a smoother loss gradient, leading to more stable training. Since the unimodal encoders remain frozen, we apply modality dropout at the input of the fusion module to prevent redundant computation. During inference, the missing modality can be handled as if the modality is dropped out, which is learned in training.

$$\mathcal{L}^{smd} = -\log p(y^i \mid, x_c^i, x_t^i, \theta) - \lambda \sum_{j \in M} \log p(y^i \mid x_j^i, \theta) \tag{1}$$

Learnable modality tokens. While conventional modality dropout effectively handled missing modalities during inference, its reliance on fixed zero matrices ($\mathbf{0}_c$ and $\mathbf{0}_t$) limited the performance of multimodal learning. Inspired by recent multimodal methods that employed learnable instructions [1,32,17], we introduce learnable modality tokens E_c and E_t to replace fixed zero matrices in our modality dropout strategy. Specifically, the unimodal predictions of the multimodal model $F_\theta(\cdot, \cdot)$, originally defined as $F_\theta(x_c^i, \mathbf{0}_t)$ and $F_\theta(\mathbf{0}_c, x_t^i,)$, are replaced with $F_\theta(x_c^i, E_t)$ and $F_\theta(E_c, x_t^i,)$, respectively. This adaptation enhances the model's generalization of missing modalities while preserving representational consistency. The modality tokens are integrated into the inputs of neural fusion modules for efficiency (shown in Fig. 1).

Contrastive multimodal fusion. Conventional contrastive learning primarily focused on unimodal representations to enhance the performance of downstream tasks [27,29,14,7,35]. In contrast, our method incorporates fused multimodal representations into contrastive learning, facilitating better cross-modal representation binding (see Fig. 1 [b]) to improve both unimodal and multimodal performance. Following [35], we employ the sigmoid-based contrastive loss for computational efficiency. We adopt supervised contrastive learning, leveraging label information, as it has been shown to outperform self-supervised approaches [35,13]. Let z_c^i, z_t^i, and z_f^i denote the encoded representations for the image, tabular, and image-tabular data of the i-th patient, respectively. For a batch of n patients with indices $N = \{1, 2, ..., n\}$, we define the contrastive loss between modalities i and j as Eq. 2, where the $a(u, v) = 1$ for positive pair ($y^u = y^v$) and $a(u, v) = -1$ for negative pair ($y^u \neq y^v$). Our approach builds upon conventional contrastive learning, initially defined as $\mathcal{L}^{con} = \mathcal{L}_{c,t}^{con}$. Unlike softmax-based contrastive learning, which requires separate directional losses between modalities (i.e., $\mathcal{L}_{c,t}^{con} \neq \mathcal{L}_{t,c}^{con}$), simoid-based contrastive learning is undirectional, meaning $\mathcal{L}_{c,t}^{con} = \mathcal{L}_{t,c}^{con}$. We extend contrastive learning to fused representations, incorporating multimodal information into the alignment process. Our final contrastive loss is defined as $\hat{\mathcal{L}}^{con} = \mathcal{L}_{c,t}^{con} + \mathcal{L}_{c,f}^{con} + \mathcal{L}_{t,f}^{con}$. This way, the multimodal representation z_f^i serves as a robust intermediary between unimodal ones z_c^i and z_t^i

to improve representation alignment and more effective multimodal fusion.

$$\mathcal{L}_{i,j}^{con} = \sum_{u \in N} \sum_{v \in N} \log \frac{1}{1 + e^{a(u,v)(-tz_i^u \cdot z_j^v + b)}}, \tag{2}$$

3 Experiment

To validate our proposed method, we conducted experiments on two publicly available datasets: the Multimodal Pulmonary Embolism (PE) dataset [38,9] and the National Lung Screening Trial (NLST) dataset [22]. The PE dataset consists of 1,837 chest CT scans from 1,794 patients, with EHRs containing PE diagnosis results. Among these CT scans, 1,111 (60.48%) are labeled as PE-positive. Following prior work [38,9], we formulated the PE detection task as a binary classification problem, utilizing both CT image and tabular data. The NLST dataset comprises 64,117 chest CT scans from 12,498 patients, each associated with EHRs. The objective, as outlined in the CT foundation model demo [33], is to predict future cancer occurrence at one-year and two-year intervals, treating these as independent classification tasks. The dataset includes 868 (6.95%) and 1,438 (11.51%) CT scans labeled as cancer-positive within one and two years, respectively. Our experimental setup aligns with these objectives, evaluating model performance in PE detection and cancer prediction by integrating multimodal learning techniques.

Experimental setting. We employed a four-fold cross-validation strategy at the patient level for both datasets. Within each fold, we further reserved 10% of the training data for validation. To benchmark our approach, we compared it against recent unimodal and multimodal baselines. Additionally, we conducted ablation studies to systematically evaluate the contributions of individual components of our proposal. Since our tasks were essentially doing classification, we utilized two categories of classification evaluation metrics: 1) Probability-estimation metrics (assess model confidence without thresholding), including area under the receiver operating characteristic curve (AUROC), average precision (AP), and the area under the risk-coverage curve (AURC). 2) decision-making metrics (evaluate binary classification with predefined threshold, e.g., 0.5), including Matthews correlation coefficient (MCC) and the F-score. We report the best and second-best results using bold and underlined fonts, respectively. During training, we used the AUROC from validation data to capture the best model, as it has been shown to be robust to class imbalance [21] For the PE detection task, we report all the evaluation metrics. In contrast, for cancer prediction tasks using NLST dataset, we focus on probability-estimation metrics for clarity and consistency, given the high sensitivity of decision-based metrics to threshold selection in extremely class-imbalanced settings.

Implementation details. For PE detection, we utilized two unimodal encoders, PENet [9] (for CT) and FT-Transformer [5] (for tabular data), followed

by a lightweight one-layer MLP (for multimodal fusion) and a linear regression head, which predicted label probabilities. To reduce the computational overhead, we froze the unimodal encoders during training, updating only the MLP and head, resulting in an efficient model with only 2.52 M *additional* training parameters. The CT images were resized to 256×256; the training batch size was set to 8. In prior PE dataset studies [9,38], models were trained using 24-slice CT windows due to computational constraints. At inference, the patient-level predictions were aggregated using maximum pooling across CT windows. However, we found that feeding the entire CT directly improved inference performance, so we adopted full-image training and inference. For the cancer prediction tasks, we employed the CT foundation model [33] as our image encoder. When we implemented the CT foundation model as an image-only baseline method, we followed the demo to train a two-layer MLP to predict the labels using embeddings extracted by the CT foundation model. The training batch size was set to 128. For all tasks, we used SHAP [20] to select the most relevant attributes in tabular data. We adopted RadFusion's attributes process and removed duplicates for the PE dataset, and collected cancer-unrelated attributes from official releases for the NLST dataset. After tuning and ablating, the top-8 (from 1226) and top-16 (from 36) attributes were selected for the PE and NLST datasets, respectively. We used AdamW [19] optimizer with 1×10^{-4} learning rate and 1×10^{-4} weight decay to train our method for 150 epochs. We simply used $\lambda = 1$ for all experiments, as it demonstrated robustness within the range 0.5 to 2.0. A fixed random seed was carefully maintained throughout training to ensure a fair comparison across experiments.

Table 1. Result of PE dataset

	Inference		AUROC↑	AP↑	AURC↓	MCC↑	F-score↑
	CT	Table					
PENet [9]	✓	-	0.758	0.609	0.475	0.207	0.538
PENet † [9]	✓	-	0.778	0.680	0.442	0.379	0.614
Ours †	✓	-	**0.801**	**0.724**	**0.422**	**0.451**	**0.647**
ElasticNet [39]	-	✓	**0.758**	**0.581**	**0.487**	**0.370**	0.564
FT-Transformer [5]	-	✓	0.745	0.539	0.510	0.351	**0.597**
Ours †	-	✓	0.751	0.558	0.499	0.352	0.594
DAFT [24]	✓	✓	0.739	0.536	0.511	0.334	0.595
DAFT † [24]	✓	✓	0.629	0.459	0.566	0.174	0.437
DAFT-64 [24]	✓	✓	0.700	0.459	0.566	0.259	0.561
DAFT-64 † [24]	✓	✓	0.616	0.479	0.560	0.139	0.434
TabAttention [6]	✓	✓	0.738	0.551	0.505	0.271	0.565
RadFusion [38]	✓	✓	0.811	0.676	0.438	0.294	0.572
RadFusion † [38]	✓	✓	0.819	0.716	0.422	0.418	0.633
RadFusion (FT)	✓	✓	0.803	0.642	0.454	0.288	0.567
RadFusion (FT) †	✓	✓	0.815	0.707	0.426	0.425	0.640
Ours †	✓	✓	**0.842**	**0.775**	**0.397**	**0.499**	**0.676**

PE dataset results. Table 1 represents the performance comparison between our method and existing approaches, including image-only models (PENet [9]), tabular-only models (ElasticNet [39] and FT-Transformer [5]), and the multimodal models (DAFT [24], TabAttention [6], and RadFusion [38]). We categorize results based on inference settings: image-only, tabular-only, and image-tabular. Our method was trained using both image and tabular data while also being robust to modality missingness during inference. The results show that our approach outperformed conventional image-only and tabular-only methods. Compared to PENnet, we increased the AUROC and AP from 0.758 to 0.801 and from 0.609 to 0.724, respectively. While ElasticNet exhibited the best tabular-only performance due to the limited number of training samples, our method achieved the best performance in the image-tabular settings. RadFusion, in its original implementation, utilized PENet and ElasticNet. To ensure a fair comparison, we also evaluated a RadFusion variant that replaced ElastiveNet with FT-Transformer (denoted as RadFusion [FT]). Our method, with only a few additional training parameters (2.52 M), outperformed Radfusion, improving image-tabular AUROC from 0.803 to 0.842 and demonstrating superior multimodal fusion efficiency and robustness. Furthermore, the additional training only cost less than 5 minutes on a Tesla v100 (16GB) GPU, making it a highly efficient plug-and-play solution.

Table 2. Results of NLST dataset

	Cancer in 2 years			Cancer in 1 year		
	AUROC↑	AP↑	AURC↓	AUROC↑	AP↑	AURC↓
CT foundation model [33]	0.729	0.070	0.956	0.700	0.050	0.972
Ours (CT only)	0.732	0.068	0.956	0.780	0.068	0.969
ElasticNet [39]	0.808	0.199	0.938	0.830	0.132	0.962
FT-Transformer [5]	0.837	0.246	0.933	0.925	**0.279**	**0.949**
Ours	**0.857**	**0.247**	**0.931**	**0.926**	**0.279**	**0.949**

NLST dataset results. Table 2 represents the evaluation results on the NLST dataset, comparing our method against the CT foundation model [33], ElasticNet [39], and FT-Transformer [5]. We report our model's performance under image-only (denoted as Ours [CT-only]) and image-tabular inference settings. Our method outperformed the CT foundation model in both two-year and one-year cancer prediction tasks, improving AUROC from 0.729 to 0.732 and 0.700 to 0.780, respectively. When incorporating tabular data, our model achieved the best overall performance, demonstrating the benefit of our multimodal learning. Interestingly, we observed that tabular data dominated the prediction in this dataset, as the performance gap between our image-tabular model and the tabular-only model was minimal. This suggested that morphometric attributes in the tabular data provide highly discriminative features for cancer prediction. Despite this, our approach successfully integrated CT imaging and tabular data,

offering the best overall performance while improving image-only one at a low additional training cost. this observation aligns with prior work [7], which demonstrated the effectiveness of morphometric attributes in contrastive learning for medical analysis.

Ablation study. To assess the contribution of individual components in our proposed method, we conducted ablation studies shown in Table 3. For simplicity, we report only AUROC, as it provides a comprehensive evaluation across all tasks. We evaluated our method across image-only and image-tabular inference settings. We denote the one-year and two-year cancer prediction tasks as NLST_1 and NLST_2, respectively. We first train the model using only the standard cross-entropy loss $\mathcal{L}^{base}$, establishing a baseline. The proposed simultaneous modality dropout $\mathcal{L}^{smd}$ slightly improved the conventional one $\mathcal{L}^{md}$. The AUROC values were improved from 0.836 to 0.840 and 0.712 to 0.722 for image-tabular PE detection and image-only NLST_2, respectively. We also noticed that the model generally converged faster when using the proposed simultaneous modality dropout, reducing the required epochs from about 300 to 50 (6 times faster). This feature may become more useful when a more costly end-to-end training or tine-tuning is used. Integrating learnable modality tokens further enhanced the performance for both $\mathcal{L}^{md}$ and $\mathcal{L}^{smd}$. Adding conventional contrastive learning $\mathcal{L}^{con}$ before target training with $\mathcal{L}^{smd}$ and learnable modality tokens further improved the performance. Finally, replacing $\mathcal{L}^{con}$ with the proposed contrastive learning $\hat{\mathcal{L}}^{con}$ consistently achieved the best performance across all tasks, regardless of the difference in scale of encoders in different experiments. These results validated the effectiveness of the proposed simultaneous modality dropout, learnable modality tokens, and contrastive multimodal fusion, demonstrating their collective contribution to robust multimodal learning. Although the overall improvements were observed, the source of the error for the misclassified patients is not clear, which requires deeper investigation.

Table 3. Ablation study results (AUROC↑)

Training loss	Modality token	Pretraining loss	Image-tabular inference			Image-only inference		
			PE	NLST_2	NLST_1	PE	NLST_2	NLST_1
$\mathcal{L}^{base}$	/	-	0.837	0.847	0.919	/	/	/
$\mathcal{L}^{md}$	-	-	0.836	0.850	0.917	0.797	0.714	0.707
$\mathcal{L}^{md}$	✓	-	0.840	0.855	0.915	0.796	0.728	0.741
$\mathcal{L}^{smd}$	-	-	0.838	0.851	**0.926**	0.800	0.716	0.738
$\mathcal{L}^{smd}$	✓	-	0.840	0.855	0.920	0.800	0.722	0.751
$\mathcal{L}^{smd}$	✓	$\mathcal{L}^{con}$	**0.842**	0.856	**0.926**	0.800	0.726	0.771
$\mathcal{L}^{smd}$	✓	$\hat{\mathcal{L}}^{con}$	**0.842**	**0.857**	**0.926**	**0.801**	**0.732**	**0.780**

4 Conclusion

We introduced a novel multimodal learning framework that integrates modality dropout with contrastive multimodal pertaining to enhance disease detection and prediction from CT images and tabular data. Our approach incorporated learnable modality tokens to improve missing modality awareness and leveraged fused multimodal representations in contrastive learning for improved alignment across modality representations. Our method offers a minimal-cost upgrade path to multimodal learning using any frozen unimodal encoder with high improvement gain. Through evaluations on three multimodal prediction tasks from two datasets, we demonstrated the effectiveness of our method in both unimodal and multimodal inference settings, showcasing its practical applicability. From a clinical application aspect, our method requires only available modalities at input, functioning seamlessly when modalities are partially available. Further improvements are anticipated if the model is trained end-to-end. Our framework is scalable to additional modalities beyond CT and EHR data. Future work will explore this scalability and investigate its applicability in conjunction with large language models, particularly the major decoder-only architectures, advancing the potential of multimodal learning in clinical applications.

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

References

1. Chen, B., et al.: A Unified Model for Longitudinal Multi-Modal Multi-View Prediction with Missingness. In: MICCAI. pp. 410–420 (2024). https://doi.org/10.1007/978-3-031-72390-2_39
2. Ding, X., et al.: HiA: Towards Chinese Multimodal LLMs for Comparative High-Resolution Joint Diagnosis. In: MICCAI. pp. 575–586 (2024)
3. Feng, Y., et al.: Unified Multi-modal Learning for Any Modality Combinations in Alzheimer's Disease Diagnosis. In: MICCAI. pp. 487–497 (2024)
4. Gao, Y., et al.: MEDBind: Unifying Language and Multimodal Medical Data Embeddings. In: MICCAI. pp. 218–228 (2024). https://doi.org/10.1007/978-3-031-72390-2_21
5. Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting deep learning models for tabular data. In: NeurIPS (2021). https://doi.org/10.1145/3704728
6. Grzeszczyk, M.K., et al.: TabAttention: Learning Attention Conditionally on Tabular Data. In: MICCAI. pp. 347–357 (2023). https://doi.org/10.1007/978-3-031-43990-2_33
7. Hager, P., et al.: Best of Both Worlds: Multimodal Contrastive Learning with Tabular and Imaging Data. In: CVPR. pp. 23924–23935 (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.02291
8. Huang, S.C., et al.: Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. Sci Rep **10**(1), 22147 (Dec 2020)

9. Huang, S.C., et al.: PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging. npj Digit. Med. **3**(1), 1–9 (Apr 2020). https://doi.org/10.1038/s41746-020-0266-y
10. Hussen Abdelaziz, A.o.: Modality Dropout for Improved Performance-driven Talking Faces. In: ICMI. pp. 378–386 (Oct 2020)
11. Jain, K., et al.: MMBCD: Multimodal Breast Cancer Detection from Mammograms with Clinical History. In: MICCAI. pp. 144–154 (2024)
12. Jiang, B., et al.: MGDR: Multi-modal Graph Disentangled Representation for Brain Disease Prediction. In: MICCAI. pp. 302–312 (2024)
13. Khosla, P., et al.: Supervised Contrastive Learning. In: NeurIPS. vol. 33, pp. 18661–18673 (2020). https://doi.org/10.5555/3495724.3497291
14. Kim, D., et al.: Learning Cross-Modal Contrastive Features for Video Domain Adaptation. In: ICCV. pp. 13598–13607 (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.01336
15. Kim, K., Lee, Y., Park, D., Eo, T., Youn, D., Lee, H., Hwang, D.: LLM-Guided Multi-modal Multiple Instance Learning for 5-Year Overall Survival Prediction of Lung Cancer. In: MICCAI. pp. 239–249 (2024)
16. Krishna, G., et al.: Modality Drop-Out for Multimodal Device Directed Speech Detection Using Verbal and Non-Verbal Features. In: ICASSP. pp. 8240–8244 (Apr 2024). https://doi.org/10.1109/ICASSP48485.2024.10446421
17. Lee, Y.L., et al.: Multimodal Prompting with Missing Modalities for Visual Recognition. In: CVPR. pp. 14943–14952 (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.01435
18. Liu, S., et al.: Multi-modal Data Fusion with Missing Data Handling for Mild Cognitive Impairment Progression Prediction. In: MICCAI. pp. 293–302 (2024)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
20. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: NeurIPS. vol. 30. Curran Associates, Inc. (2017)
21. McDermott, M., et al.: A closer look at auroc and auprc under class imbalance. In: NeurIPS. vol. 37, pp. 44102–44163 (2024)
22. National Lung Screening Trial Research Team: Data from the National Lung Screening Trial (NLST). The Cancer Imaging Archive (2013)
23. Neverova, N., et al.: ModDrop: Adaptive Multi-Modal Gesture Recognition. IEEE TPAMI **38**(8), 1692–1706 (Aug 2016). https://doi.org/10.1109/TPAMI.2015.2461544
24. Pölsterl, S., et al.: Combining 3D Image and Tabular Data via the Dynamic Affine Feature Map Transform. In: MICCAI. pp. 688–698 (2021). https://doi.org/10.1007/978-3-030-87240-3_66
25. Qi, A., et al.: Multimodal Emotion Recognition with Vision-language Prompting and Modality Dropout. In: Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing. pp. 49–53 (Oct 2024)
26. Qiu, S., et al.: Multimodal deep learning for Alzheimer's disease dementia assessment. Nat Commun **13**(1), 3404 (Jun 2022). https://doi.org/10.1038/s41467-022-31037-5
27. Qu, L., et al.: Multi-modal Data Binding for Survival Analysis Modeling with Incomplete Data and Annotations. In: MICCAI. pp. 501–510 (2024)
28. Radford, A., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: ICML (Jul 2021), https://proceedings.mlr.press/v139/radford21a.html

29. Robinet, L., et al.: DRIM: Learning Disentangled Representations from Incomplete Multimodal Healthcare Data. In: MICCAI. pp. 163–173 (2024). https://doi.org/10.1007/978-3-031-72384-1_16
30. Xiong, C., et al.: MoME: Mixture of Multimodal Experts for Cancer Survival Prediction. In: MICCAI. pp. 318–328 (2024)
31. Xiong, Z., et al.: Multi-modality 3D CNN Transformer for Assisting Clinical Decision in Intracerebral Hemorrhage. In: MICCAI. pp. 522–531 (2024)
32. Xu, J., et al.: Temporal Neighboring Multi-modal Transformer with Missingness-Aware Prompt for Hepatocellular Carcinoma Prediction. In: MICCAI. pp. 79–88 (2024). https://doi.org/10.1007/978-3-031-72378-0_8
33. Yang, L., et al.: Advancing multimodal medical capabilities of gemini. arXiv preprint arXiv:2405.03162 (2024)
34. Yu, J., et al.: Coca: Contrastive captioners are image-text foundation models. TMLR (2022), https://openreview.net/forum?id=Ee277P3AYC
35. Zhai, X., et al.: Sigmoid Loss for Language Image Pre-Training. In: ICCV. pp. 11941–11952 (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.01100
36. Zhang, S., Du, S., Sun, C., Li, B., Shao, L., Zhang, L., Wang, K., Liu, Z., Tian, J.: M2Fusion: Multi-time Multimodal Fusion for Prediction of Pathological Complete Response in Breast Cancer. In: MICCAI. pp. 458–468 (2024)
37. Zhou, Q., et al.: PathM3: A Multimodal Multi-task Multiple Instance Learning Framework for Whole Slide Image Classification and Captioning. In: MICCAI. pp. 373–383 (2024). https://doi.org/10.1007/978-3-031-72083-3_35
38. Zhou, Y., et al.: RadFusion: Benchmarking Performance and Fairness for Multimodal Pulmonary Embolism Detection from CT and EHR (Nov 2021). https://doi.org/10.48550/arXiv.2111.11665
39. Zou, H., Hastie, T.: Regularization and Variable Selection Via the Elastic Net. J. R. Stat. Soc. Ser. B Methodol. **67**(2), 301–320 (Apr 2005). https://doi.org/10.1111/j.1467-9868.2005.00503.x