ClinLoop: Human-AI Collaborative Agents for Multimodal Breast Cancer Diagnosis under Incomplete Clinical Data

Hui Xie[1,2⋆], Ziyang Wang[3,*], Sheng Zhu[2], Zhiqiang Wang[2], Zhuoneng Zhang[1], Fangyijie Wang[4], Tingfeng Zhang[5], Jie Ma[5], and Tao Tan[1] (✉)

[1] Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR, 999078, China
[2] Department of Radiotherapy, Affiliated Hospital (Clinical College) of Xiangnan University, Chenzhou, 423000, China
[3] School of Computer Science and Digital Technologies, Aston University, Birmingham, B47ET, UK
[4] School of Medicine, University College Dublin, Dublin, Ireland
[5] Department of Radiology, Shenzhen People's Hospital, The Second Clinical Medical College of Jinan University, Shenzhen, 518020, China

taotanjs@gmail.com

Abstract. Breast cancer imaging is multimodal but often incomplete: in routine care, patients may have CT, MRI, or digital mammography, and only a subset has all modalities. We study diagnosis from heterogeneous, partially observed imaging and propose ClinLoop, a human–AI workflow with two agents. A diagnosis agent predicts malignancy from the available modalities using modality-specific encoders and a missing-modality fusion module. A quality-assurance agent estimates uncertainty with an ensemble and defers low-confidence or discordant cases. Deferred cases are reviewed by a clinician, and the confirmed labels are fed back to update the model. This loop reduces manual review while keeping a safe fallback for uncertain and out-of-distribution inputs. We evaluate on a private cohort of 3,108 patients with non-uniform modality availability (CT+MRI, CT-only, MRI-only, or mammography-only) and TNM staging. Experimental results demonstrate that ClinLoop improves diagnosis performance and efficiency compared with single-modality models and standard multimodal fusion baselines, with consistent gains across modality subgroups. ClinLoop supports practical deployment under missing data and limited expert time.

Keywords: Human–AI Collaboration · Multi-agent Systems · Multimodal Imaging · Missing Modalities · Breast Cancer

1 Introduction

Deep learning has become a core technique for medical image analysis, delivering strong performance in detection, classification, and segmentation tasks

⋆ These authors contributed equally to this work.

across modalities [12, 17, 21]. Despite this progress, clinical deployment remains constrained by three practical factors: limited expert labels, heterogeneous acquisition protocols, and the need for reliable decisions that can be audited and safely integrated into clinical workflows [8,10,16]. Recent work [20] on human–AI collaboration and agentic decision-making has highlighted a pragmatic direction: instead of forcing full automation, models should know when to defer [2,13], route uncertain cases to experts [14], and use feedback to improve over time [15]. In this setting, an agent is not simply a predictor, but a workflow component that can estimate confidence, trigger quality control, and coordinate human review under time constraints.

Breast cancer diagnosis is a representative and high-impact example where these constraints are pronounced. Clinicians routinely rely on multiple imaging modalities, such as CT, MRI, and digital mammography, to capture complementary anatomical and pathological information [3, 4]. However, real-world cohorts are rarely complete: patients may have only one modality due to cost, contraindications, scheduling, or pathway differences, while a minority have multiple modalities [7]. This operational missingness breaks a common assumption in multimodal learning, where fusion models are trained and tested with all modalities present. Although missing-modality strategies (e.g., modality dropout [5], imputation [19], or mixture-of-experts fusion [11]) can improve robustness, they do not directly address a key clinical requirement: when modality evidence is insufficient or out-of-distribution, the system should explicitly flag uncertainty and involve a clinician rather than producing overconfident errors.

We propose ClinLoop, a human–AI collaborative loop designed for multimodal breast cancer diagnosis under incomplete clinical data. ClinLoop contains two cooperative agents. The first agent performs diagnosis from the available modalities using modality-specific encoders and a missing-modality fusion module, enabling prediction for CT-only, MRI-only, mammography-only, and multimodality patients. The second agent performs quality assurance by estimating uncertainty with an ensemble and deferring low-confidence or internally inconsistent cases. For deferred cases, a clinician reviews the evidence and provides a confirmed label, which is then fed back to update the model. This design aligns with clinical practice: the system automates straightforward cases, reserves expert time for high-risk cases, and maintains a traceable decision path for safety and accountability.

Our main contributions are: (i) A private, real-world breast imaging cohort with heterogeneous modality availability across CT, MRI, and digital mammography, reflecting operational missingness and accompanied by clinical staging information. (ii) A two-agent collaborative framework for diagnosis and quality assurance that supports selective deferral and clinician review, enabling label-efficient learning and safer deployment. (iii) A multimodal learning design that operates with arbitrary subsets of modalities at inference time, improving robustness compared with standard fusion pipelines that assume complete inputs. (iv) Extensive evaluation showing improved discrimination and calibration over single-modality and multimodal baselines, and stronger performance

Table 1. Modality availability by patient.

Group	n	%
CT + MRI	844	27.156
CT only	520	16.731
MRI only	1,315	42.310
DM (X-ray) only	429	13.803
Total	3,108	100.00

than general-purpose multimodal large-model assistants (e.g., ChatGPT- [1] and Qwen-style [18] systems used as decision aids), while substantially reducing required clinician review time through uncertainty-based triage.

2 Dataset and Problem Definition

2.1 Clinical Dataset

We study a real-world, multimodal breast imaging cohort collected from routine clinical pathways. The cohort contains N=3,108 patients with heterogeneous modality availability across CT, breast MRI, and digital mammography (DM). As commonly observed in practice, modality completeness is not guaranteed: only a small subset has paired CT and MRI, while many patients have a single modality due to pathway differences, contraindications, cost, or scheduling. Table 1 summarises the resulting operational missingness, and Fig. 1 shows the distribution of available modalities across TNM stages.

For each patient, imaging is paired with structured clinical staging information in TNM format, recorded at diagnosis and/or pre-treatment. All data collection and use were approved by the relevant institutional ethics committee(s). All images and associated metadata were de-identified before analysis. To comply with the double-blind review policy, identifying information (including institution name and any potentially traceable details) is withheld and the dataset is presented in anonymised form.

2.2 Problem Statement

Let $\mathcal{M} = \{\mathrm{CT}, \mathrm{MRI}, \mathrm{DM}\}$ denote the set of modalities. For patient i, only a subset $\mathcal{S}_i \subseteq \mathcal{M}$ is observed, yielding inputs $X_i = \{x_i^{(m)} \mid m \in \mathcal{S}_i\}$. The goal is to perform TNM staging from incomplete multimodal evidence. We model the target as $y_i = (y_i^T, y_i^N, y_i^M)$, where y_i^T, y_i^N, and y_i^M are categorical labels for tumour, nodal, and metastasis status, respectively. Accordingly, the learning objective is a multi-task classification function f that maps any available subset of modalities to calibrated predictive probabilities $p_i = f(X_i)$, without requiring imputation or test-time access to missing modalities.

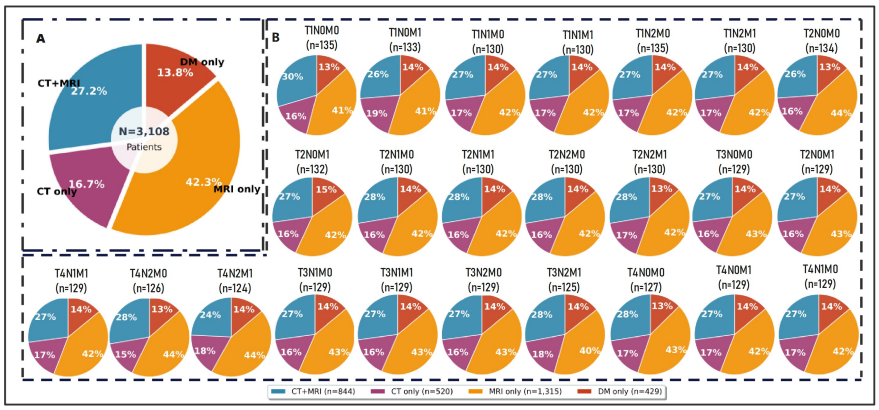

Fig. 1. Distribution of Imaging Modalities by TNM Stage. (A) Overall Distribution (N=3,108); (B) TNM Stage-Specific Distributions.

In addition to accuracy, the clinical setting requires safe operation under uncertainty. We therefore consider a selective prediction setting: the system outputs either a staging prediction $\hat{y}_i$ (auto-accept) or defers the case for clinician review (abstain). A quality-assurance decision rule g uses uncertainty and consistency signals to determine whether to accept or defer: $a_i = g(p_i, X_i) \in \{0, 1\}$, where a_i=1 indicates auto-accept and a_i=0 indicates deferral. Performance is assessed in terms of staging accuracy and calibration for accepted cases, together with deferral rate (or equivalently, clinician workload) under incomplete modality availability.

3 Methods

3.1 Overall Architecture

ClinLoop is an iterative human–AI workflow built around a multimodal diagnostic model and two supporting agents (Fig. 2). For patient i, only a subset of modalities $\mathcal{S}_i \subseteq \{\mathrm{CT}, \mathrm{MRI}, \mathrm{DM}\}$ is available, with inputs $X_i = \{x_i^{(m)} \mid m \in \mathcal{S}_i\}$. Each modality is encoded with a modality-specific backbone E_m to obtain embeddings $h_i^{(m)} = E_m(x_i^{(m)})$. A missing-modality fusion module $F(\cdot)$ aggregates the available embeddings using a modality mask:

$$z_i = F\Big(\{h_i^{(m)}\}_{m \in \mathcal{S}_i}, \mathbf{1}_{\mathcal{S}_i}\Big). \tag{1}$$

Multi-task heads predict TNM components (or a combined stage label):

$$p_i^T, p_i^N, p_i^M = G(z_i), \tag{2}$$

and the supervised objective is the sum of cross-entropies over available ground-truth labels. During training, modality-dropout is applied so the fusion learns to operate with arbitrary modality subsets.

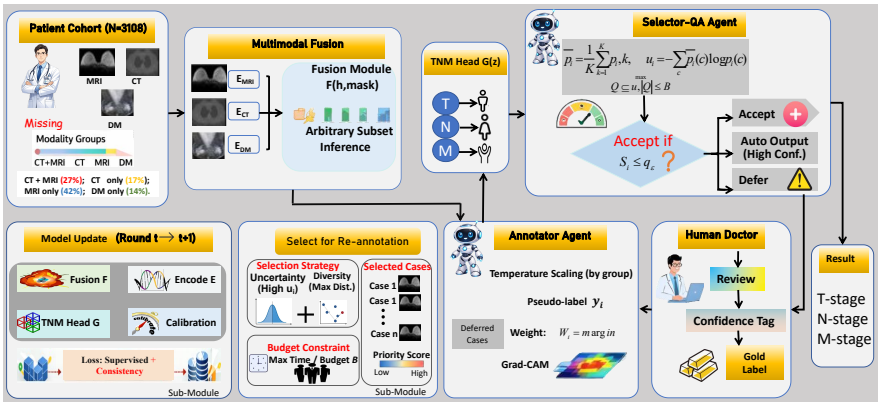

Fig. 2. ClinLoop Collaborative Multi-Agent Workflow for Label-Efficient Multimodal Breast Cancer Diagnosis under Incomplete Data.

ClinLoop runs in rounds. The Selector–QA agent determines which cases can be auto-accepted and which should be deferred for human review, and also selects additional cases for annotation under a time budget. The Annotator agent provides calibrated pseudo-labels and explanations to reduce clinician effort. The clinician reviews deferred cases and a budgeted subset of high-impact samples, producing gold labels that are fed back to update the model in the next round.

3.2 Selector–QA Agent

Uncertainty is estimated using an ensemble of K models and test-time augmentation (TTA). For sample i, we compute the mean predictive distribution $\bar{p}_i$ and an uncertainty score u_i (predictive entropy):

$$\bar{p}_i = \frac{1}{K}\sum_{k=1}^{K} p_{i,k}, \quad u_i = -\sum_{c} \bar{p}_i(c)\log \bar{p}_i(c). \tag{3}$$

To avoid selecting redundant cases, we also compute embedding similarity using fused features z_i. Given an unlabeled pool $\mathcal{U}$ and budget B, the agent selects a set $\mathcal{Q}$ by a diversity-aware objective (facility-location):

$$\max_{\mathcal{Q}\subseteq\mathcal{U},\,|\mathcal{Q}|\le B} \sum_{i\in\mathcal{Q}} u_i \;+\; \lambda \sum_{j\in\mathcal{U}} \max_{i\in\mathcal{Q}} \mathrm{sim}(z_i, z_j), \tag{4}$$

optimised with a greedy selection strategy.

For quality assurance, the agent uses conformal calibration on a held-out calibration split to derive an acceptance threshold for each task. Let $s_i = 1 - \max_c \bar{p}_i(c)$ be a nonconformity score. From calibration scores, we obtain a quantile q_ϵ. A prediction is auto-accepted only if it is sufficiently reliable:

$$\mathrm{accept}(i) = \mathbb{I}[\, s_i \le q_\epsilon \,] \cdot \mathbb{I}[\,\text{no modality-consistency violation}\,], \tag{5}$$

otherwise the case is deferred to the clinician. A simple consistency check flags cases where different ensemble members (or different modality subsets, when available) disagree on the predicted TNM component.

3.3 Annotator Agent

For selected cases, the Annotator agent generates pseudo-labels with calibration and a confidence weight. We apply modality-group temperature scaling (grouped by available modality subset) to logits before averaging ensemble predictions:

$$\tilde{p}_i = \frac{1}{K}\sum_{k=1}^{K} \mathrm{softmax}\left(\frac{\ell_{i,k}}{T_{g(i)}}\right), \quad \tilde{y}_i = \arg\max_c \tilde{p}_i(c). \tag{6}$$

Pseudo-labels are only used when they pass the same conformal reliability rule as above; otherwise they are routed for human review. For accepted pseudo-labels, we set a weight based on the prediction margin:

$$w_i = \max\big(0,\ \tilde{p}_i(c_1) - \tilde{p}_i(c_2)\big), \tag{7}$$

where c_1 and c_2 are the top-1 and top-2 classes. These weighted pseudo-labels are mixed with gold labels in training.

To support rapid verification, the agent produces a compact explanation: Grad-CAM key-slice summaries for CT/MRI and saliency maps for DM, stored alongside the decision record for auditing.

3.4 Human-in-the-Loop Doctor

The clinician reviews (i) all deferred cases from the QA gate and (ii) a budgeted subset of additional samples selected for maximal benefit under limited time. We prioritise review by an impact score that combines uncertainty and expected error:

$$r_i = u_i \cdot (1 - \max_c \bar{p}_i(c)), \tag{8}$$

and select the top cases until the time budget is reached. The clinician provides TNM labels (or stage labels) with a confidence tag (high/medium/low). High-confidence clinician labels are added as gold labels; lower-confidence labels can be added with reduced weight. After each round, the diagnostic model and calibration thresholds are updated using the expanded gold set, completing the loop.

4 Experiments & Results

4.1 Implementation Details

We evaluate on the private clinical cohort described in Sec. 2. Within each training fold, a small split is reserved for calibration (for conformal gating and temperature scaling). CT/MRI volumes and DM images are preprocessed with standard

clinical pipelines (resampling/normalisation and resizing to a fixed input size); when volumetric inputs are used, we follow a slice-based strategy and aggregate slice features with attention. The diagnostic model uses modality-specific encoders with a missing-modality fusion module conditioned on the modality mask, and multi-task heads for (T, N, M) (or a combined stage label, depending on the experiment).

Selector–QA uses an ensemble (K models with different seeds) and test-time augmentation to estimate uncertainty. Conformal thresholds are computed on the calibration split and used to auto-accept or defer predictions. In each round, the Selector–QA agent selects additional cases under a fixed budget using uncertainty + diversity, while the Annotator agent produces calibrated pseudo-labels and explanations for candidate cases. The clinician reviews all deferred cases and an optional budgeted subset of selected cases; reviewed labels are added to the training set for the next round.

4.2 Metrics

We report macro-F1 and accuracy for each TNM component (and exact-match accuracy when predicting the joint label). To quantify reliability, we report expected calibration error (ECE) [6] and Brier score [9]. To measure human workload and selective safety, we report (i) *coverage* (fraction auto-accepted), (ii) *selective risk* (error on accepted cases), and (iii) the accuracy/F1 at fixed review budgets (e.g., top-x% deferred or reviewed). Results are also stratified by modality availability group (CT+MRI, CT-only, MRI-only, DM-only).

4.3 Quantitative Results

Baselines. We compare against (1) single-modality models trained/evaluated on their respective subgroups, (2) multimodal fusion baselines that operate under missing modalities (e.g., late fusion, mixture-of-experts, modality dropout), (3) traditional classifiers on engineered features (e.g., radiomics + logistic regression/SVM), and (4) off-the-shelf multimodal LLM assistants (e.g., ChatGPT- [1] and Qwen-style [18] systems) used as black-box decision aids by providing the same case inputs (representative slices/projections and minimal clinical context) and mapping outputs to TNM labels.

Results. Table 2 reports overall TNM performance and calibration. ClinLoop achieves the best macro-F1/accuracy and improves reliability (lower ECE/Brier) compared with single-modality and missing-modality fusion baselines. The gains are consistent across modality-availability subgroups, indicating robustness to operational missingness. Off-the-shelf multimodal LLM baselines underperform on this task, with weaker calibration and less consistent staging predictions, highlighting the benefit of domain-trained models with explicit uncertainty control.

Human attention budget. Table 3 shows performance versus clinician review budget (or reviewed images/cases). Performance improves monotonically with more review, but ClinLoop reaches strong operating points with limited

Table 2. Overall TNM staging results. ↑, ↓ denote higher or lower is better.

Method	T-F1↑	N-F1↑	M-F1↑	Joint-Acc↑	ECE↓
Radiomics + LR	0.612	0.553	0.701	0.523	0.152
Radiomics + SVM	0.643	0.591	0.721	0.552	0.141
Single-modality (CT)	0.701	0.652	0.781	0.621	0.101
Single-modality (MRI)	0.772	0.713	0.802	0.671	0.082
Single-modality (DM)	0.672	0.612	0.643	0.571	0.131
Missing-modality Fusion	0.792	0.732	0.813	0.691	0.071
ChatGPT-style assistant	0.541	0.492	0.591	0.441	0.191
Qwen-style assistant	0.572	0.521	0.623	0.472	0.172
ClinLoop (Ours)	**0.851**	**0.812**	**0.873**	**0.762**	**0.036**

Table 3. Performance vs clinician review budget. Budget denotes the percentage of patient cases reviewed by clinicians.

Budget	Coverage↑	Selective Risk↓	Macro-F1↑	ECE↓
0%	0.721	0.179	0.721	0.086
5%	0.781	0.152	0.763	0.066
10%	0.823	0.123	0.802	0.051
20%	0.852	0.092	0.831	0.041
30%	0.881	0.073	0.852	0.036

review because the QA agent concentrates human effort on uncertain cases. This produces a favorable accuracy–workload trade-off, reducing required clinician attention while maintaining safety through deferral.

5 Conclusion & Discussion

We presented ClinLoop, a human–AI collaborative workflow for multimodal breast cancer staging under real-world missing modalities. By combining a diagnostic model that operates on arbitrary modality subsets with a Selector–QA agent and an Annotator agent, ClinLoop improves accuracy and calibration while reducing the amount of clinician attention required. Experiments on a private cohort with heterogeneous CT/MRI/DM availability show consistent gains over single-modality, missing-modality fusion baselines, and general-purpose multimodal LLM assistants, with a favourable performance–workload trade-off. Future work will extend ClinLoop to broader multi-centre validation, richer clinical context and downstream tasks.

Acknowledgments. This work was supported by the Science and Technology Development Fund of Macao under Grant (0041/2023/RIB2).

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

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alves, J.V., Leitão, D., Jesus, S., Sampaio, M.O.P., Liébana, J., Saleiro, P., Figueiredo, M.A.T., Bizarro, P.: A benchmarking framework and dataset for learning to defer in human-AI decision-making. Scientific Data **12**(1), 506 (apr 2025)
3. Comstock, C.E., Gatsonis, C., Newstead, G.M., Snyder, B.S., Gareen, I.F., Bergin, J.T., Rahbar, H., Sung, J.S., Jacobs, C., Harvey, J.A., Nicholson, M.H., Ward, R.C., Holt, J., Prather, A., Miller, K.D., Schnall, M.D., Kuhl, C.K.: Comparison of abbreviated breast mri vs digital breast tomosynthesis for breast cancer detection among women with dense breasts undergoing screening. JAMA **323**(8), 746–756 (02 2020)
4. Dagar, N., Sarin, M., Yadav, P., Thidwar, R.: Radiologic imaging and its current importance in breast cancer management. Frontiers in Radiology **5** (2026)
5. Gu, Y., Saito, K., Ma, J.: Learning contrastive multimodal fusion with improved modality dropout for disease detection and prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 280–290. Springer (2025)
6. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
7. Hassan, M.M., Tahsin, A., Alam, M.G.R., Alzamil, D., Garg, S., Uddin, M.Z., Choudhury, N., Fortino, G.: Explainable multimodal fusion for breast carcinoma diagnosis: A systematic review, open problems, and future directions. Computer Methods and Programs in Biomedicine **274**, 109152 (2026)
8. Hognon, C., Conze, P.H., Bourbonne, V., Gallinato, O., Colin, T., Jaouen, V., Visvikis, D.: Contrastive image adaptation for acquisition shift reduction in medical imaging. Artificial Intelligence in Medicine **148**, 102747 (2024)
9. Huang, Y., Li, W., Macheret, F., Gabriel, R.A., Ohno-Machado, L.: A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association **27**(4), 621–633 (02 2020)
10. Lekadir, K., Frangi, A.F., Porras, A.R., Glocker, B., Cintas, C., Langlotz, C.P., Weicken, E., et al.: Future-ai: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ **388** (2025)
11. Li, S., Chen, C., Han, J.: Simmlm: A simple framework for multi-modal learning with missing modality. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24068–24077 (2025)
12. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
13. Mozannar, H., Lang, H., Wei, D., Sattigeri, P., Das, S., Sontag, D.: Who should predict? exact algorithms for learning to defer to humans. In: International conference on artificial intelligence and statistics. pp. 10520–10545. PMLR (2023)
14. popat, r., ive, j.: Embracing the uncertainty in human-machine collaboration to support clinical decision-making for mental health conditions. Frontiers in Digital Health **volume 5 - 2023** (2023)
15. Vasey, B., Nagendran, M., Campbell, B., Clifton, D.A., Collins, G.S., Denaxas, S., Denniston, A.K., et al.: Reporting guideline for the early-stage clinical evaluation

of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine **28**(5), 924–933 (may 2022)
16. Wang, S., Li, C., Wang, R., Liu, Z., Wang, M., Tan, H., Wu, Y., Liu, X., Sun, H., Yang, R., Liu, X., Chen, J., Zhou, H., Ben Ayed, I., Zheng, H.: Annotation-efficient deep learning for automatic medical image segmentation. Nature Communications **12**(1), 5915 (oct 2021)
17. Wang, Z., Yang, C.: Mixsegnet: Fusing multiple mixed-supervisory signals with multiple views of networks for mixed-supervised medical image segmentation. Engineering Applications of Artificial Intelligence **133**, 108059 (2024)
18. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
19. Yang, Y., Chen, H., Chang, Z., Xiang, Y., Ye, C., Ma, T.: Incomplete learning of multi-modal connectome for brain disorder diagnosis via modal-mixup and deep supervision. In: Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 227, pp. 1006–1018. PMLR (10–12 Jul 2024)
20. Zhang, S., Yu, J., Xu, X., Yin, C., Lu, Y., Yao, B., Tory, M., Padilla, L.M., Caterino, J., Zhang, P., Wang, D.: Rethinking human-ai collaboration in complex medical decision making: A case study in sepsis diagnosis. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. CHI '24, Association for Computing Machinery, New York, NY, USA (2024)
21. Zhang, S., Metaxas, D.: On the challenges and perspectives of foundation models for medical image analysis. Medical Image Analysis **91**, 102996 (2024)