BUThink: Aligning Multiparametric Breast Ultrasound with Clinical Expert Logic via Rule-Constrained GRPO

Chenqi Xu[1,*], Ang Lv[2,*], Yanbo Liu[2], Yong Liu[1], Ting Liu[2], Guanghao Sun[1], Longfei Cong[2], Rui Hou[1,†], and Xing An[2,†]

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
rui.hou@bupt.edu.cn
[2] Shenzhen Mindray BioMedical Electronics, Co., Ltd, Shenzhen, China
anxing@mindray.com
∗ Equal contribution. † Corresponding authors.

Abstract. Accurate breast ultrasound diagnosis benefits from integrating multiple complementary image types. Training multimodal large language models (MLLMs) for this setting is challenging because breast ultrasound reports rarely contain step-by-step diagnostic reasoning, and collecting such rationales at scale requires substantial expert effort. Here we present BUThink, a three-stage framework for guideline-based reasoning while reducing reliance on manually written long-form reasoning annotations. First, we automatically convert clinical guidelines into executable logical rules and use them to synthesize stepwise diagnostic rationales, forming chain-of-thought (CoT) style supervision. Second, we introduce an adaptive prompting scheme for supervised fine-tuning that conditions the model on multiparametric ultrasound inputs, maps modality-specific observations to clinical questions, and resolves conflicting cues by prioritizing higher-risk interpretations. Third, we apply Group Relative Policy Optimization (GRPO) with rule-based rewards to refine diagnostic accuracy and report quality without relying on human preference labels. We train BUThink on 49,145 images, and evaluate it on a private cohort and the public dataset US3M. This framework achieves state-of-the-art performance on US3M, with 92.34% accuracy and 95.65% recall. By replacing manual long-form rationale annotation with guideline-derived supervision and rule-guided optimization, BUThink provides a practical approach for building more reliable medical MLLMs.

Keywords: Multimodal Large Language Models · Group Relative Policy Optimization · Breast Ultrasound Diagnosis · Chain-of-Thought.

1 Introduction

Medical image interpretation can be time-consuming and knowledge-intensive, motivating the development of computer-aided diagnosis (CAD) systems to support clinical decision making. CAD has been explored for decades in breast

imaging and related applications, aiming to improve consistency and assist with screening and diagnosis [10, 18, 22, 13]. However, many existing CAD systems primarily output scores or labels with limited transparency, making it difficult for clinicians to understand model rationale and calibrate trust in safety-critical use. Recent multimodal large language models (MLLMs) offer a potential path toward decision support that combines visual evidence with natural-language justification [22, 5, 9, 17]. Prior studies have improved ultrasound report generation through supervised fine-tuning, modality-aware visual encoding, and retrieval-augmented generation (RAG)-based corpus expansion [8, 16, 20], but practical limitations remain, including incomplete diagnostic outputs, constraints on multi-image input, and repetitive template-like text.

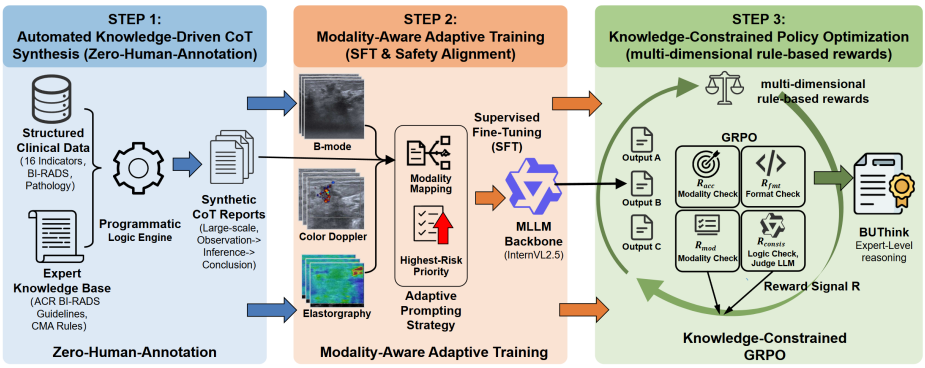

Fig. 1. Overall architecture of BUThink. The model is trained in three stages: (1) Automated Knowledge-Driven CoT Synthesis, (2) Modality-Aware Adaptive Training, and (3) Knowledge-Constrained Policy Optimization.

A major limitation is the form of available text supervision. Routine ultrasound reports are usually brief and emphasize final impressions, whereas guideline-based diagnosis depends on explicit criteria and standardized descriptors, such as the ACR BI-RADS lexicon in breast imaging [2]. As a result, many datasets provide labels or short reports but lack step-by-step rationales linking imaging findings to assessments, which limits how diagnostic MLLMs can learn and be evaluated [5, 9]. This problem is more pronounced in multiparametric ultrasound, where B-mode, Doppler, and elastography may provide partially inconsistent evidence. Without explicit guidance for reconciling such evidence under clinical rules and risk-sensitive decision making, models may generate unstable rationales or overconfident conclusions.

Guided by this motivation, we propose BUThink, a three-stage training framework for multiparametric ultrasound diagnosis. BUThink derives reasoning supervision by translating guideline criteria into executable rules and generating stepwise CoT traces, avoiding the need to collect long-form clinician-written rationales at scale. These traces support supervised fine-tuning with adaptive

prompts that map modality-specific observations to clinical questions and apply a highest-risk priority rule when evidence conflicts across images or modalities [12]. Finally, GRPO with rule-based rewards aligns the output with diagnostic, structural, modality-coverage, and consistency constraints without human preference annotations [14]. We evaluate BUThink on a large private multimodal ultrasound dataset and the public US3M benchmark [19], with ablations used to examine each training component.

2 Method

BUThink is a three-stage training framework for generating guideline-consistent breast ultrasound reports from multiparametric ultrasound inputs. During both training and inference, the model receives only the unified diagnostic prompt and multiparametric ultrasound images as input; ground-truth clinical indicators are not provided to the model as input. Given a multiparametric ultrasound study, the model produces a structured report with a BI-RADS assessment and supporting reasoning. Training proceeds as follows. First, an automated rule engine converts standardized clinical indicators into synthetic step-by-step reasoning traces. Second, modality-aware adaptive prompting maps each ultrasound image type to the corresponding clinical assessment target and applies a highest-risk priority rule when observations conflict. Third, GRPO uses rule-based rewards to optimize output format, BI-RADS correctness, modality coverage, and consistency between the reasoning and final answer.

This design reduces the need for clinicians to write long-form reasoning annotations: expert-provided observation labels are used offline to synthesize CoT targets and compute supervised training loss, while the model itself learns to infer these observations from images. The same rule base is used both to synthesize training traces and to score candidate outputs during optimization, providing a practical training signal when reasoning-rich text supervision is limited.

2.1 Automated Knowledge-Driven CoT Synthesis

Limited access to detailed diagnostic reasoning text and the cost of long-form expert annotation present challenges for training medical MLLMs. To address this limitation, we develop an automated system for generating reasoning traces. Instead of asking clinicians to write long explanations, we translate the ACR BI-RADS lexicon and CMA diagnostic guidelines into logical rules, allowing standard clinical indicators to be converted into structured CoT supervision.

For each patient, the system evaluates 16 clinical indicators, including lesion characteristics, blood-flow features, and surrounding tissue context. From these inputs, it generates a reasoning trace that follows guideline logic, highlights higher-risk signs, and assigns a BI-RADS category as the final assessment. Rule-based generation enables scalable dataset construction while keeping the synthetic reasoning aligned with the underlying guideline criteria. Because clinically grounded ordering is essential for coherent diagnostic reasoning, we further

construct a shuffled-step variant by rearranging the steps within the reasoning trace into a clinically incoherent sequence while preserving the same content. This design allows us to isolate the contribution of reasoning order.

To determine an effective organization for synthetic reasoning traces, we define three clinically motivated structures. **Indicator-wise reasoning** evaluates all 16 indicators sequentially before aggregating them into a BI-RADS category. **Region-wise reasoning** groups observations into three anatomical regions, namely lesion, lymph node, and glandular background, performs region-level assessment, and then derives the final BI-RADS category by integrating the regional findings. **Malignancy-sign-only reasoning** considers only suspicious indicators associated with malignancy while omitting benign findings. These variants are evaluated in Section 3.2.

2.2 Modality-Aware Adaptive Prompting Strategy

Multiparametric ultrasound introduces additional complexity because each patient may have multiple image types and several images per modality. We therefore use a structured prompting strategy during supervised fine-tuning. First, the prompt asks the model to identify the image type, such as B-mode or Color Doppler, and associate it with the corresponding clinical assessment target. B-mode images are used mainly for lesion morphology, whereas Doppler images support blood-flow assessment. Second, when observations conflict within or across modalities, the prompt applies a highest-risk priority rule: suspicious findings take precedence over benign-appearing cues. This rule is designed to reduce missed high-risk findings under ambiguous evidence.

The prompt also guides the model to generate reasoning in a structured CoT format. In the final inference workflow, the MLLM first identifies the 16 image-based clinical indicators from the input multiparametric ultrasound images while screening for major suspicious signs, such as architectural distortion or abnormal lymph nodes. It then uses these indicators to construct a stepwise reasoning trace and predicts the patient-level BI-RADS category together with the benign or malignant risk statement. The reasoning trace is written in a dedicated `<think>` block, and the final BI-RADS category with the risk statement is written in an `<answer>` block. This structure encourages explicit organization of evidence across modalities before the final conclusion.

2.3 Aligning Clinical Fidelity via Knowledge-Constrained GRPO

After supervised fine-tuning, we apply GRPO as the final training stage to align generated reports with clinical and structural constraints. This stage optimizes candidate outputs using rewards derived from guideline rules and report-format requirements.

GRPO avoids training a separate reward model, which typically requires large-scale human preference labels. Instead, candidate outputs are scored with rule-based criteria derived from clinical guidelines. A general-purpose large language model is used only for the consistency check between the reasoning trace

and final answer. The overall reward for candidate i is defined as a weighted combination of four components:

$$R_i = \alpha R_{\text{acc}}^{(i)} + \beta R_{\text{fmt}}^{(i)} + \gamma R_{\text{mod}}^{(i)} + \delta R_{\text{cons}}^{(i)}, \qquad \alpha + \beta + \gamma + \delta = 1. \tag{1}$$

To model relative preference within the same input group, rewards are normalized into group-wise advantages:

$$A_i = \frac{R_i - \mu_R}{\sigma_R}, \qquad \mu_R = \frac{1}{N}\sum_{j=1}^{N} R_j, \quad \sigma_R = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(R_j - \mu_R)^2 + \epsilon}. \tag{2}$$

The policy is then optimized with a clipped GRPO objective and KL regularization:

$$\begin{aligned} \max_{\theta} \ \mathcal{L}(\theta) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \min\Big(\rho_i(\theta)A_i,\ \text{clip}\big(\rho_i(\theta), 1-\varepsilon, 1+\varepsilon\big)A_i\Big)\right] \\ - \lambda\, \mathbb{E}[D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})]\,, \\ \rho_i(\theta) = \frac{\pi_\theta(y_i \mid x)}{\pi_{\theta_{\text{old}}}(y_i \mid x)}. \end{aligned} \tag{3}$$

This formulation increases the probability of candidates with higher rewards while constraining policy drift from the reference model. The reward components are defined as follows.

Accuracy Reward (R_{acc}). This term evaluates whether the predicted BI-RADS category and risk level match the reference labels.

Format Reward (R_{fmt}). This term enforces the required output structure. The reasoning must appear inside `<think>` tags and the conclusion inside `<answer>` tags. The reward is 1 if the output follows the format exactly and 0 otherwise. This term is designed primarily to ensure parseable and complete report generation rather than to directly optimize the final diagnostic endpoint.

Modality Reward (R_{mod}). This term encourages coverage of key evidence in the reasoning trace. It checks whether the reasoning mentions all 16 predefined observation indicators and whether the final output states both the BI-RADS category and pathological risk. Each correctly included indicator contributes a fractional score, yielding a higher reward for more complete evidence coverage.

Consistency Reward (R_{cons}). This term measures whether the final conclusion is supported by the preceding reasoning. We use Qwen3 as an automatic judge to assess whether the `<answer>` content is consistent with the `<think>` rationale [4, 3]. The reward is 1 when the conclusion is supported by the reasoning and 0.5 when it is inconsistent or insufficiently supported. We use this soft penalty because the deterministic accuracy reward already supervises endpoint correctness, and a nonzero score reduces the risk that ambiguous wording judged by an LLM overrides clinical labels.

The weights are set to $\alpha = 0.5$, $\beta = 0.2$, $\gamma = 0.2$, and $\delta = 0.1$, satisfying $\alpha + \beta + \gamma + \delta = 1$. They are chosen to reflect the clinical importance and

reliability of each reward component. Accuracy receives the largest weight ($\alpha = 0.5$) because correct BI-RADS prediction is the primary optimization objective. Format ($\beta = 0.2$) and modality ($\gamma = 0.2$) encourage structured output and comprehensive evidence coverage, while consistency receives the smallest weight ($\delta = 0.1$) because it is evaluated by an LLM judge rather than deterministic clinical rules.

3 Experiments

Datasets. We evaluate BUThink on two datasets: (1) a large-scale private multimodal clinical dataset used for training and ablation, and (2) the public US3M dataset used for comparison with prior work. ***Private Multimodal Clinical Dataset.*** Our primary dataset is a private multicenter repository collected from 32 tertiary hospitals across multiple regions, with patient ages ranging from 18 to 90 years. It contains 5,012 breast ultrasound patients, including 3,220 benign and 1,792 malignant diagnoses. Each patient record includes 9 to 12 images from multiple scan planes and modalities, such as B-mode and Doppler, totaling 49,145 images. Malignancy ground truth is established by pathological biopsy. For each patient, senior physicians from tertiary hospitals provided annotations for 16 clinical indicators covering lesion morphology, lymph-node status, and glandular context, together with the standardized ACR BI-RADS category. These annotations are used to construct CoT supervision and training targets, but are not used as model inputs during either training or inference. The dataset was split at the patient level into training, validation, and internal test sets at an 8:1:1 ratio, with 4,010 patients for training, 501 for validation, and 501 for independent testing. ***Public Benchmark Dataset (US3M).*** We also evaluate on US3M, which contains 1,532 multimodal images from 248 patients collected at the Second Affiliated Hospital of Harbin Medical University. The dataset provides binary diagnostic labels, where 0 denotes benign and 1 denotes malignant tumors. We use US3M to compare BUThink with previously reported state-of-the-art methods, including TDF-Net.

Evaluation Metrics. We quantify the model's ability to identify pathological status using standard clinical metrics. Following the US3M protocol, we report accuracy (ACC), precision, recall (sensitivity), and F1-score. Because conventional text-generation metrics such as BLEU and ROUGE correlate weakly with expert assessment of medical reasoning [23], they are not used as primary endpoints in the main text. The Supplementary Material provides a compact diagnostic prompt summary, indicator-level recognition summaries, a qualitative example, and an adapted MedThink-Bench-style scoring-point analysis of reasoning coverage.

Implementation Details. We fine-tune InternVL-2.5 using the proposed training pipeline on four NVIDIA A100 GPUs. The batch size is 2 per GPU, with gradient accumulation over 2 steps, giving an effective batch size of 8. We use AdamW with an initial learning rate of 1×10^{-4} and a warmup ratio of 0.05. In the GRPO stage, the group size is set to $G = 4$, and the KL coefficient is

set to $\lambda = 0.04$ to balance policy updates and training stability. The pipeline is implemented with the Swift library.

3.1 Comparisons with State-of-the-Art Methods

Table 1 reports performance on US3M. BUThink outperforms all compared baselines, including multimodal fusion networks and recent transformer-based approaches. BUThink achieves 92.34% accuracy and a 93.24% F1-score, improving over TDF-Net by 1.32 and 2.48 percentage points, respectively. It also reaches 95.65% recall, 4.48 percentage points higher than TDF-Net (91.17%) and higher than TFormer (87.96%) and MAL (84.17%).

Higher recall is clinically important for breast ultrasound classification because it reflects sensitivity to patients with malignant findings. Because several compared methods do not release training code or checkpoints, Table 1 uses the results reported in the corresponding original papers under the public US3M protocol. Differences in training data, preprocessing, and implementation may therefore affect strict comparability. The results indicate that guideline-based CoT supervision and rule-constrained GRPO improve both diagnostic performance and output control.

3.2 Main Results and Ablation Analysis

Table 1. Performance on US3M.

Method	ACC	Prec.	Rec.	F1
MSMFN [11]	83.67	83.91	84.97	83.54
ResGRU [1]	82.85	82.81	81.95	82.06
FusionM4Net [15]	84.49	84.63	85.66	84.34
MAL [7]	85.30	85.84	84.17	84.59
MMGLF [6]	84.08	83.87	84.23	83.70
TFormer [21]	86.12	87.21	87.96	85.99
TDF-Net [19]	91.02	91.21	91.17	90.76
BUThink (ours)	**92.34**	**91.03**	**95.65**	**93.24**

Table 2. Ablation results.

Method	ACC	Prec.	Rec.	F1
w/o GRPO	94.17	96.01	95.21	95.61
w/o CoT	86.11	90.60	88.33	89.45
w/o Format Reward	95.97	97.27	96.67	96.97
w/o Accuracy Reward	90.00	94.74	90.00	92.31
w/o Modality Reward	93.75	95.03	95.63	95.33
w/o Consistency Reward	93.19	94.50	95.11	94.80
Indicator-wise CoT	90.28	94.37	90.83	92.57
Malignancy-sign-only CoT	85.42	92.52	85.00	88.60
Shuffled-step CoT	70.28	77.59	77.92	77.75
Region-wise CoT (baseline)	96.25	97.89	96.46	97.17

We further evaluate BUThink on the private multimodal clinical dataset and conduct ablation studies to assess each component. As shown in Table 2, the full framework achieves 96.25% accuracy and a 97.17% F1-score. The w/o GRPO variant keeps CoT-supervised fine-tuning but removes the final GRPO stage, reducing accuracy to 94.17%. The w/o CoT variant bypasses Stage 1 CoT synthesis and uses label-only supervision before GRPO, causing a larger drop to 86.11% accuracy. These results indicate that guideline-derived reasoning traces and GRPO each contribute to the final performance.

The reward ablations show that the diagnostic accuracy reward contributes most strongly: removing it lowers accuracy from 96.25% to 90.00%. Removing the format reward has a smaller effect, with accuracy remaining at 95.97%,

which suggests that this reward mainly preserves complete and parseable report structure after the model has already learned the required output format, rather than directly serving the final diagnostic endpoint.

Removing the modality reward lowers accuracy to 93.75% and the F1-score to 95.33%, consistent with reduced coverage of evidence across modalities. Removing the consistency reward lowers accuracy to 93.19% and the F1-score to 94.80%, suggesting that agreement between the reasoning trace and final answer helps stabilize report generation. Together, these ablations indicate that both CoT synthesis and multi-term rule-based rewards contribute to the final performance.

Among the three CoT structures, the region-wise design performs best, with 96.25% accuracy, compared with 90.28% for indicator-wise reasoning and 85.42% for malignancy-sign-only reasoning. This result suggests that organizing reasoning by anatomical regions better matches clinical diagnostic workflows for multiparametric ultrasound. The shuffled-step variant achieves only 70.28% accuracy and a 77.75% F1-score, showing that preserving the same reasoning content is not sufficient when the clinical order is disrupted.

Qualitative reasoning. Qualitative inspection further shows that BUThink first identifies image-based clinical indicators from multiparametric ultrasound images, organizes suspicious findings into a stepwise reasoning trace, and then predicts the patient-level BI-RADS category with the benign or malignant judgment. A representative example is provided in the Supplementary Material.

4 Conclusion

In this work, we present BUThink, a three-stage training framework for automated breast ultrasound interpretation. BUThink reduces reliance on clinician-written long-form reasoning annotations by converting BI-RADS guidelines into an automated rule engine that generates explicit CoT reasoning traces for supervised training. Modality-aware fine-tuning supports consistent use of multiparametric inputs, and GRPO further aligns outputs with rule-based diagnostic and structural constraints. Overall, the results show that combining guideline-derived reasoning synthesis with constraint-based optimization is a practical strategy for medical multimodal models when reasoning-rich text supervision is limited.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant number: 82402374).

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

References

1. Ahmad, M., Bajwa, U.I., Mehmood, Y., Anwar, M.W.: Lightweight resgru: a deep learning-based prediction of sars-cov-2 (covid-19) and its severity classification using multimodal chest radiography images. Neural Computing and Applications **35**(13), 9637–9655 (2023). https://doi.org/10.1007/s00521-023-08200-0

2. American College of Radiology: ACR BI-RADS Atlas: Breast Imaging Reporting and Data System. American College of Radiology, Reston, VA, 5th edn. (2013), includes Mammography, Ultrasound, and Magnetic Resonance Imaging
3. Arias-Duart, A., Martin-Torres, P.A., Hinjos, D., Bernabeu-Perez, P., Ganzabal, L.U., Mallo, M.G., Gururajan, A.K., Lopez-Cuena, E., Alvarez-Napagao, S., Garcia-Gasulla, D.: Automatic evaluation of healthcare llms beyond question-answering. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers). pp. 108–130. Association for Computational Linguistics, Albuquerque, New Mexico (2025). https://doi.org/10.18653/v1/2025.naacl-short.10, https://aclanthology.org/2025.naacl-short.10/
4. Croxford, E., Gao, L., First, M., Pellegrino, F., Schnier, C., Caskey, R.: Evaluating clinical ai summaries with large language models as judges. npj Digital Medicine **8**, 640 (2025). https://doi.org/10.1038/s41746-025-02005-2
5. Guo, X., Chai, W., Li, S.Y., Wang, G.: Llava-ultra: Large chinese language and vision assistant for ultrasound. In: Proceedings of the 32nd ACM international conference on multimedia. pp. 8845–8854 (2024). https://doi.org/10.1145/3664647.3681584
6. Jia, N., Jia, T., Zhao, L., Ma, B., Zhu, Z.: Multi-modal global- and local-feature interaction with attention-based mechanism for diagnosis of alzheimer's disease. Biomedical Signal Processing and Control **95**, 106404 (2024). https://doi.org/10.1016/j.bspc.2024.106404
7. Jiao, M., Liu, H., Liu, J., Ouyang, H., Wang, X., Jiang, L., Yuan, H., Qian, Y.: Mal: Multi-modal attention learning for tumor diagnosis based on bipartite graph and multiple branches. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 175–185. Springer (2022). https://doi.org/10.1007/978-3-031-16437-8_17
8. Li, J., Su, T., Zhao, B., Lv, F., Wang, Q., Navab, N., Hu, Y., Jiang, Z.: Ultrasound report generation with cross-modality feature alignment via unsupervised guidance. IEEE Transactions on Medical Imaging **44**(1), 19–30 (2025). https://doi.org/10.1109/TMI.2024.3424978
9. Li, Z., Li, M., Wang, W., Huang, Q.: Ultrasound report generation with fuzzy knowledge and multi-modal large language model. Expert Systems with Applications **292**, 128555 (2025). https://doi.org/10.1016/j.eswa.2025.128555
10. Lu, S.Y., Wang, S.H., Zhang, Y.D.: Safnet: A deep spatial attention network with classifier fusion for breast cancer detection. Computers in Biology and Medicine **148**, 105812 (2022). https://doi.org/10.1016/j.compbiomed.2022.105812
11. Meng, Z., Zhu, Y., Pang, W., Tian, J., Nie, F., Wang, K.: Msmfn: an ultrasound based multi-step modality fusion network for identifying the histologic subtypes of metastatic cervical lymphadenopathy. IEEE Transactions on Medical Imaging **42**(4), 996–1008 (2023). https://doi.org/10.1109/TMI.2022.3222541
12. Qian, X., Pei, J., Zheng, H., Xie, X., Yan, L., Zhang, H., Han, C., Gao, X., Zhang, H., Zheng, W., et al.: Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. Nature biomedical engineering **5**(6), 522–532 (2021). https://doi.org/10.1038/s41551-021-00711-2
13. Sexauer, R., Hejduk, P., Borkowski, K., Ruppert, C., Weikert, T., Dellas, S., Schmidt, N.: Diagnostic accuracy of automated acr bi-rads breast density classification using deep convolutional neural networks. European Radiology **33**(7), 4589–4596 (2023). https://doi.org/10.1007/s00330-023-09474-7

14. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024), https://arxiv.org/abs/2402.03300
15. Tang, P., Yan, X., Nan, Y., Xiang, S., Krammer, S., Lasser, T.: Fusionm4net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification. Medical Image Analysis **76**, 102307 (2022). https://doi.org/10.1016/j.media.2021.102307
16. Wang, X., Li, Y., Wang, F., Wang, S., Li, C., Jiang, B.: R2gencsr: Mining contextual and residual information for llms-based radiology report generation. arXiv preprint arXiv:2408.09743 (2024), https://arxiv.org/abs/2408.09743
17. Wang, Z., Sun, Y., Li, Z., Yang, X., Chen, F., Liao, H.: Llm-rg4: Flexible and factual radiology report generation across diverse input contexts. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8250–8258 (2025). https://doi.org/10.1609/aaai.v39i8.32890
18. Xing, J., Chen, C., Lu, Q., Cai, X., Yu, A., Xu, Y., Xia, X., Sun, Y., Xiao, J., Huang, L.: Using bi-rads stratifications as auxiliary information for breast masses classification in ultrasound images. IEEE Journal of Biomedical and Health Informatics **25**(6), 2058–2070 (2021). https://doi.org/10.1109/JBHI.2020.3034804
19. Yan, P., Gong, W., Li, M., Zhang, J., Li, X., Jiang, Y., Luo, H., Zhou, H.: Tdf-net: Trusted dynamic feature fusion network for breast cancer diagnosis using incomplete multimodal ultrasound. Information Fusion **112**, 102592 (2024). https://doi.org/10.1016/j.inffus.2024.102592
20. Yang, Y., You, X., Zhang, K., Fu, Z., Wang, X., Ding, J., Sun, J., Yu, Z., Huang, Q., Han, W., et al.: Spatio-temporal and retrieval-augmented modelling for chest x-ray report generation. IEEE Transactions on Medical Imaging **44**(7), 2892–2905 (2025). https://doi.org/10.1109/TMI.2025.3554498
21. Zhang, Y., Xie, F., Chen, J.: Tformer: A throughout fusion transformer for multi-modal skin lesion diagnosis. Computers in biology and medicine **157**, 106712 (2023). https://doi.org/10.1016/j.compbiomed.2023.106712
22. Zhao, Z., Wang, S., Gu, J., Zhu, Y., Mei, L., Zhuang, Z., Cui, Z., Wang, Q., Shen, D.: Chatcad+: Toward a universal and reliable interactive cad using llms. IEEE Transactions on Medical Imaging **43**(11), 3755–3766 (2024). https://doi.org/10.1109/TMI.2024.3398350
23. Zhou, S., Xie, W., Li, J., et al.: Automating expert-level medical reasoning evaluation of large language models. npj Digital Medicine **9**(1), 34 (2025). https://doi.org/10.1038/s41746-025-02208-7