MediFact at MEDIQA-CORE-Task-2 2026: Interaction-Aware Multimodal Fusion for Fine-Grained Radiology Report Discrepancy Analysis

Nadia Saeed[1,*]

[1]Department of Computer Sciences, Quaid-i-Azam University, Islamabad, Pakistan

Abstract

Radiology report discrepancy analysis requires fine-grained reasoning over medical images and clinical text. The ImageCLEFmed MEDIQA-CORE Task 2 challenge focuses on evaluating discrepancies between preliminary radiology reports and candidate edits using abdominal CT volumes. This work proposes an interaction-aware multimodal framework combining DeBERTa-v3 contextual text embeddings with 3D CNN-based CT feature extraction for agreement prediction, severity assessment, and edit-type classification. To improve discrepancy reasoning, semantic relationships between reports and candidate edits were explicitly modeled using difference and element-wise interaction features. The framework further incorporates robust CT preprocessing, modality-balanced fusion, layer normalization, and weighted multitask optimization for stable multimodal learning. Among 28 participating teams, the proposed MediFact system achieved second place in the official challenge evaluation under the Natural setting, obtaining a composite score of 0.58 compared with the organizer's baseline score of 0.48. The proposed framework further improved agreement prediction from 0.63 to 0.76 and severity assessment from 0.61 to 0.76, demonstrating stronger multimodal discrepancy reasoning capability. The implementation of our MediFact framework will be publicly available[1].

Keywords

Multimodal Deep Learning, Medical Image Analysis, DeBERTa-v3 Transformer, Multitask Learning Framework

1. Introduction

Radiology reporting is an iterative clinical workflow in which preliminary reports generated by junior radiologists are subsequently reviewed and refined by experienced clinicians. Proposed report modifications may correct diagnostic inaccuracies, incorporate omitted findings, or clarify ambiguous interpretations. Accurate discrepancy analysis is therefore essential for improving reporting consistency, reducing diagnostic errors, and strengthening quality assurance in radiology practice [1, 2].

Recent advances in multimodal artificial intelligence have demonstrated promising capabilities in jointly reasoning over medical images and clinical text for tasks such as report generation, disease classification, and clinical decision support [3, 4]. However, automated radiology discrepancy analysis is challenging. Clinically important edits are often subtle. They also depend on the clinical context. Accurate alignment between medical images and report text is essential [5].

RADAR is a recently introduced multimodal benchmark designed to evaluate 3D image-based radiology report review using clinically realistic trainee-to-attending radiology workflows [6]. RADAR is constructed from expert-annotated abdominal CT examinations containing real-world report revisions. The benchmark evaluates model reasoning across three clinically important dimensions. Image-grounded agreement assessment. Discrepancy severity estimation. And edit type classification, including correction, addition, and clarification. Baseline experiments using large foundation models such as Gemini-3-Pro and Qwen3.5-plus demonstrate strong performance on edit type prediction [7, 8]. While highlighting substantial limitations in achieving robust overall discrepancy reasoning

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

drnadiasaeed@qau.edu.pk (N. Saeed)

https://github.com/NadiaSaeed (N. Saeed)

0000-0003-0394-4691 (N. Saeed)

performance. These findings indicate that current approaches are not sufficient. Better multimodal reasoning is needed to detect subtle differences between radiology reports and imaging findings.

The ImageCLEFmed MEDIQA-CORE 2026 Task 2 challenge (conducted by ImageCLEF 2026 [9]) addresses this problem through a multimodal learning framework involving abdominal CT volumes, preliminary radiology reports, and candidate report edits [10]. The task requires models to reason over volumetric images and clinical text. It determines whether a proposed edit agrees with the CT findings. It also assesses the severity of the discrepancy and classifies the edit category. Unlike conventional vision-language classification tasks, this task requires a fine-grained understanding of how candidate edits change the clinical meaning of the original report.

Previous studies have investigated report generation, factual consistency verification, hallucination detection, and discrepancy identification in radiology reports. Most existing methods focus on sentence-level consistency or synthetic error detection. Fine-grained edit-level reasoning based on clinically meaningful report revisions remains comparatively underexplored.

Multitask learning frameworks have been widely adopted in clinical artificial intelligence. They jointly optimize related prediction objectives using shared feature representations. Such approaches often improve model generalization, learning efficiency, and robustness. They achieve this by leveraging task-level correlations across clinically related outputs [11, 12]. In multimodal medical applications, multitask optimization has shown benefits. It jointly learns diagnostic, semantic, and contextual representations from heterogeneous clinical data sources [13, 14].

Existing multimodal approaches frequently rely on direct concatenation of independently extracted visual and textual embeddings [3]. This approach is effective for coarse-grained prediction tasks. However, it may inadequately capture subtle semantic interactions between original reports and candidate edits. This is particularly true when clinically significant discrepancies arise from minor textual modifications. Furthermore, many existing frameworks lack explicit interaction modeling mechanisms. These mechanisms are needed to represent nuanced report-edit relationships in clinically complex settings.

To address these limitations, this work proposes an interaction-aware multimodal fusion framework for radiology discrepancy analysis. Fig. 1 illustrates the overall architecture of the proposed MediFact framework.

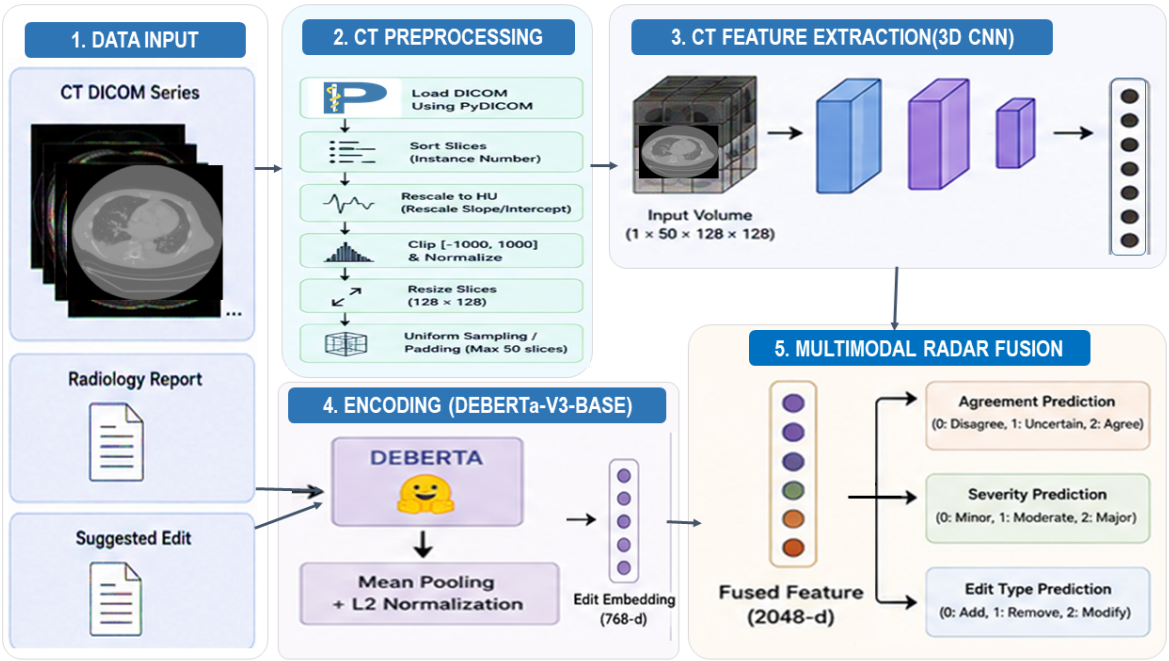

Figure 1: Overview of the proposed MediFact framework for multimodal radiology report assessment. Text and CT image features are fused for agreement, severity, and edit type prediction.

Our approach uses contextual text representations generated by the DeBERTa-v3 transformer architecture [15]. It combines these representations with volumetric CT features extracted through a lightweight 3D convolutional neural network [16]. The proposed framework explicitly models semantic relationships between reports and candidate edits. It uses difference-aware and element-wise interaction representations to improve discrepancy reasoning. The system also incorporates robust CT preprocessing, modality-balanced multimodal fusion, and stabilized multitask optimization. These components improve generalization across clinically correlated prediction tasks [17, 12]. Experimental evaluation on the MEDIQA-CORE 2026 Task 2 benchmark demonstrates the effectiveness of the proposed framework. The framework achieved second place overall on the challenge leaderboard.

2. Methodology

2.1. Overview of the Proposed Framework

The proposed framework is designed to jointly analyze radiology reports, suggested report edits, and corresponding CT scans for the ImageCLEFmed MEDIQA 2026 Task 02 challenge [10]. The framework integrates transformer-based textual representations with volumetric CT image features through a multimodal fusion strategy referred to as RADAR [6]. The overall pipeline consists of CT image preprocessing, textual and visual feature extraction, multimodal fusion, multitask learning, and inference.

2.2. CT Image Preprocessing

The dataset contains multiple CT scan series for each case, including ABD_PEL, COR_BODY, SAG_BODY, and SCOUT views. Different single-series and multi-series combinations were experimentally evaluated during model development. Among all tested configurations, the ABD_PEL series achieved the best performance. Therefore, we selected it for the final framework. The selection is clinically meaningful. The ABD_PEL series contains the most complete axial anatomical information for abdominal and pelvic findings. These findings are highly relevant to the challenge objective. In contrast, SCOUT images mainly provide low-detail localization information. COR_BODY and SAG_BODY views introduce additional variability and computational complexity. They do not provide significant performance improvement. CT scans were processed directly from DICOM files using the `pydicom` library [18]. Valid slices containing pixel data were loaded and sorted according to their `InstanceNumber` to preserve anatomical consistency throughout the volume. Pixel intensities were converted into Hounsfield Units (HU) using the DICOM rescale parameters using Eq. 1.

$$I_{HU} = I_{raw} \times S + B \tag{1}$$

where S and B denote the rescale slope and intercept, respectively.

To suppress intensity outliers and improve stability across scans, voxel values were clipped to the range $[-1000, 1000]$:

$$I_{clip} = \text{clip}(I_{HU}, -1000, 1000) \tag{2}$$

The clipped values were normalized using fixed HU scaling:

$$I_{norm} = \frac{I_{HU} + 1000}{2000} \tag{3}$$

All slices were resized to 128×128 resolution using anti-aliased interpolation, and all CT volumes were standardized to a fixed depth of 50 slices through uniform sampling or zero-padding. Finally, all processed volumes were converted to `float32` format for computational efficiency and stable GPU processing.

2.3. Text Representation Learning

The radiology report and suggested edit text were encoded separately using the pretrained DeBERTa-v3-base transformer model developed by Microsoft [15]. The model was selected because of its strong contextual language understanding capability and effectiveness in generating semantic representations from clinical text.

Each text sequence was tokenized and truncated to a maximum length of 256 tokens before being passed through the transformer encoder. Instead of relying only on the CLS token representation, mean pooling was applied across the final hidden states to generate dense semantic embeddings. The resulting embeddings were normalized using L2 normalization. This improved feature consistency and stability during multimodal fusion. Missing or empty text entries were represented using zero-valued vectors. This ensured robustness throughout the pipeline.

2.4. CT Feature Extraction

A lightweight 3D convolutional neural network (3D CNN) extracted volumetric features from the CT scans. The network consists of stacked 3D convolution layers. It uses ReLU activation and max-pooling operations. It also includes adaptive average pooling and a fully connected projection layer [16].

This lightweight architecture enables efficient learning of spatial and contextual information from volumetric CT data while maintaining low computational complexity. The extracted CT representation was projected into a 512-dimensional embedding space. The generated CT embeddings were normalized prior to multimodal fusion to stabilize representation learning.

2.5. Interaction-Aware Multimodal Fusion

To better capture semantic relationships between the original report and the proposed edit, interaction-aware textual features were introduced.

Given the report embedding r and edit embedding e, two interaction representations were computed:

$$d = r - e, \quad p = r \odot e \tag{4}$$

where d represents directional semantic difference and p denotes element-wise semantic interaction between both textual representations.

The final textual feature representation was constructed as:

$$t = [r \mid e \mid d \mid p] \tag{5}$$

The CT scaling factor(α) was evaluated using two model variants (0.2 and 0.3), whose results are reported in the paper. The setting of 0.3 consistently achieved better overall composite performance and was therefore selected for the final model.

$$x = [t \mid \alpha \cdot c] \tag{6}$$

A reduced weighting factor was applied to the CT representation to prevent visual feature dominance and maintain balanced learning between textual and imaging modalities.

2.6. RADAR Multitask Fusion Network

The fused multimodal representation was processed using the proposed RADAR fusion framework. Before feature learning, Layer Normalization was applied to stabilize feature distributions across modalities. The normalized feature vector was then passed through a shared fully connected feature learning backbone consisting of dense layers, ReLU activations, and dropout regularization:

$$h = \mathrm{Dropout}(\mathrm{ReLU}(W_2(\mathrm{Dropout}(\mathrm{ReLU}(W_1 x))))) \tag{7}$$

The learned shared representation was forwarded to three independent task-specific classification heads for agreement prediction, severity classification, and edit type classification:

$$\hat{y}_a, \hat{y}_s, \hat{y}_e = f(h) \tag{8}$$

This multitask learning strategy enables the framework to learn shared semantic relationships across related clinical tasks while preserving task-specific discriminative information.

2.7. Model Training Strategy

The multimodal fusion network was optimized using the Adam optimizer with a learning rate of 1×10^{-4}. Several hyperparameter configurations were evaluated during development, and the final settings were selected based on best validation performance. Cross-entropy loss was independently computed for agreement, severity, and edit classification tasks:

$$\mathcal{L}_a = CE(\hat{y}_a, y_a), \quad \mathcal{L}_s = CE(\hat{y}_s, y_s), \quad \mathcal{L}_e = CE(\hat{y}_e, y_e) \tag{9}$$

The final multitask objective function was defined as:

$$\mathcal{L} = w_a\mathcal{L}_a + w_s\mathcal{L}_s + w_e\mathcal{L}_e \tag{10}$$

The task-specific loss weights(w_a, w_c, w_e) were determined through validation experiments, with higher emphasis on agreement prediction due to its stronger influence on the final evaluation metric. To improve optimization stability and prevent exploding gradients during backpropagation, gradient clipping was applied:

$$\|g\| \leq 1.0 \tag{11}$$

The proposed MediFact model was trained for five epochs. It used mini-batch optimization with a batch size of 8. Five epochs were sufficient for convergence. This was confirmed using validation performance.

2.8. Inference and Prediction Pipeline

During inference, the trained MediFact model was evaluated in inference mode. This disabled dropout behavior and stabilized prediction consistency. Inference was performed in a no-gradient execution context. This reduced memory consumption and computational overhead.

The report embedding, edit embedding, and CT embedding were extracted independently. They were normalized before fusion. All feature vectors were converted to `float32` format. This maintained datatype consistency across modalities.

3. Results and Discussion

The proposed MediFact framework was evaluated using the official evaluation protocol provided by the ImageCLEFmed MEDIQA 2026 Task 02 organizers.

Table 1
Performance comparison on the ImageCLEFmed MEDIQA 2026 Task 02 benchmark using the official organizer evaluation metrics. The official evaluation code provided by the organizers was used for all reported results [2].

Team	Composite (Mixed)	Agreement (Mixed)	Severity (Mixed)	Edit Type (Mixed)	Composite (Natural)	Agreement (Natural)	Severity (Natural)	Edit Type (Natural)
MEDIQA Organizers	**0.39**	**0.68**	**0.56**	**0.82**	0.48	0.63	0.61	**0.84**
anubhav27	0.33	0.52	0.52	0.68	0.58	**0.86**	0.76	0.72
MediFact (Proposed)	0.32	0.43	0.51	0.70	**0.58**	0.76	**0.76**	0.72
ClinicalVLM-TW	0.22	0.53	0.38	0.63	0.29	0.49	0.46	0.65
krishnatewari	0.05	0.22	0.51	0.12	0.08	0.24	0.76	0.14

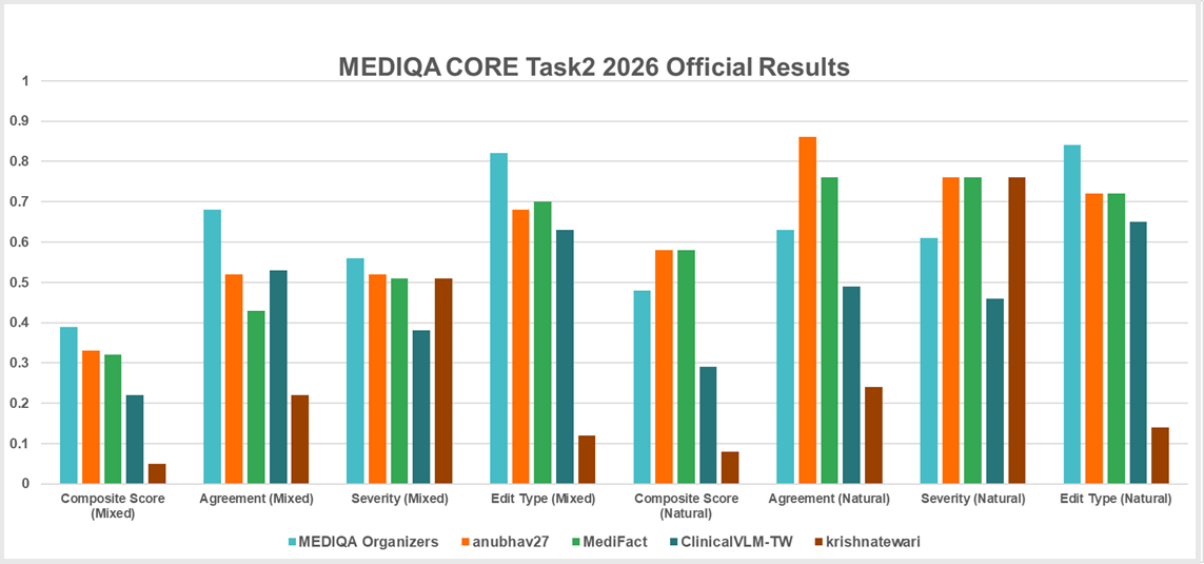

Figure 2: Performance comparison between the proposed MediFact framework and competing systems on the ImageCLEFmed MEDIQA 2026 Task 02 benchmark using the official evaluation metrics under Mixed and Natural evaluation settings.

As illustrated in Fig. 2, the proposed MediFact framework achieved strong overall performance. It performed particularly well under the Natural evaluation setting. The results demonstrate the effectiveness of the proposed MediFact framework for multimodal clinical edit assessment. Under the Natural evaluation setting, MediFact achieved the highest composite score of 0.58. It outperformed the organizer baseline (0.48) and other participating systems. The framework also achieved a score of 0.76 for agreement prediction. It achieved a score of 0.76 for severity classification. Overall, the results validate the effectiveness of the proposed RADAR framework. The framework performs agreement prediction, severity assessment, and edit type classification for clinical report correction tasks.

To analyze the effectiveness of the proposed multimodal framework, two different fusion strategies were experimentally evaluated. As illustrated in Fig. 3, the proposed interaction-aware RADAR fusion framework achieved improved overall performance compared with the baseline multimodal fusion strategy.

The first configuration, referred to as the *Baseline Multimodal Fusion* approach, relied on direct concatenation of report embeddings, edit embeddings, and CT image features. The second configuration, referred to as the *Interaction-Aware RADAR Fusion* approach, introduced interaction-based textual representations together with weighted CT feature integration and enhanced multitask optimization.

Table 2
Comparison between the baseline multimodal fusion framework and the proposed interaction-aware RADAR fusion framework using the official ImageCLEFmed MEDIQA 2026 evaluation metrics.

Method	Composite (Mixed)	Agreement (Mixed)	Severity (Mixed)	Edit Type (Mixed)	Composite (Natural)	Agreement (Natural)	Severity (Natural)	Edit Type (Natural)
Baseline Multimodal Fusion	0.29	**0.48**	0.51	0.70	0.49	0.72	0.76	0.72
Interaction-Aware RADAR Fusion (Proposed)	**0.32**	0.43	0.51	0.70	**0.58**	**0.76**	0.76	0.72

As shown in Fig. 4, the proposed interaction-aware RADAR fusion framework consistently improved overall composite performance compared with the baseline multimodal fusion approach. The results demonstrate that the proposed interaction-aware RADAR fusion framework achieved improved overall performance compared with the baseline multimodal fusion approach. In particular, the proposed framework obtained the highest Natural composite score of 0.58, improving upon the baseline score of 0.49. A noticeable improvement was also observed in agreement prediction performance, where the score increased from 0.72 to 0.76.

The performance gain can be attributed to the introduction of interaction-aware textual representa-

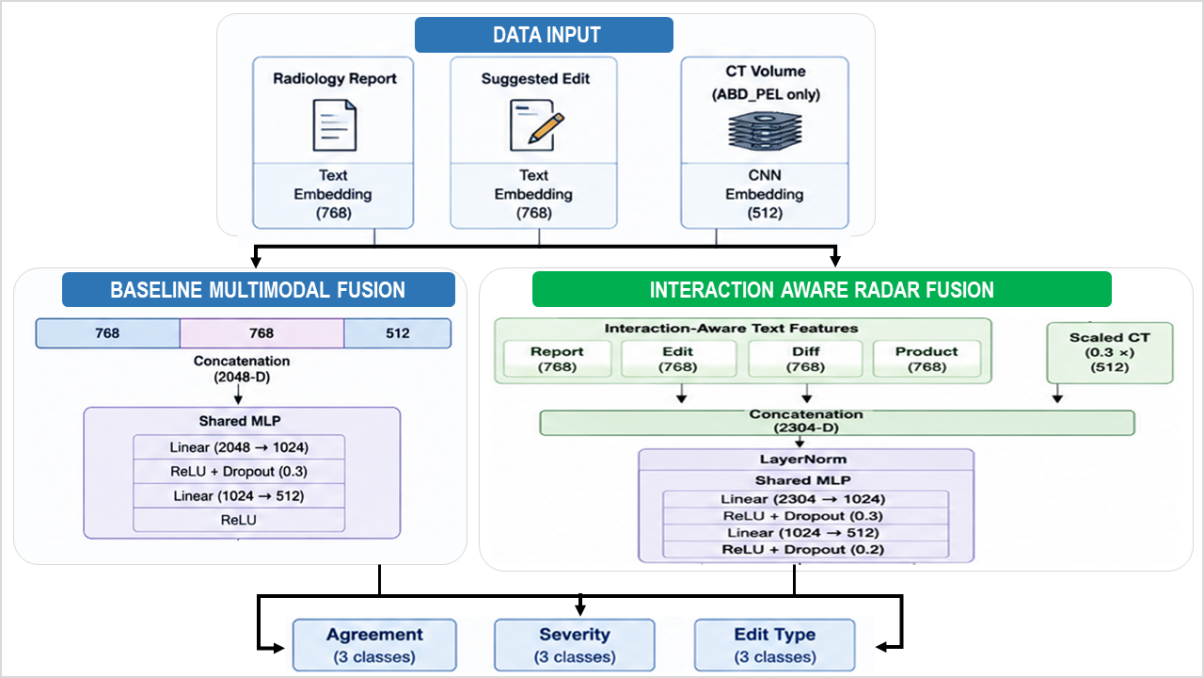

Figure 3: Performance comparison between the baseline multimodal fusion approach and the proposed interaction-aware RADAR fusion framework under Mixed and Natural evaluation settings. The proposed framework achieves improved composite performance and stronger agreement prediction through interaction-aware feature learning and balanced multimodal fusion.

tions. Although the baseline approach achieved slightly higher agreement performance in the Mixed setting, the proposed framework demonstrated substantially stronger generalization in the Natural setting, which more closely reflects real-world clinical reporting scenarios. This suggests that the interaction-aware fusion strategy improves robustness against natural report variability and enhances clinically meaningful semantic understanding.

Overall, the experimental results confirm that the proposed interaction-aware RADAR fusion framework provides a more expressive and balanced multimodal representation compared with standard feature concatenation approaches, leading to improved performance for multimodal radiology report correction and assessment tasks.

Acknowledgements

Thank to the organizers of the ImageCLEFmed MEDIQA 2026 Task 02 challenge for providing the dataset, evaluation framework, and benchmarking platform used in this work.

Declaration on Generative AI

During the preparation of this work, used freely available AI tools for language refinement and illustrative figure design assistance. All reviewed and verified by the author.

References

[1] N. Kadom, T. S. Cook, M. A. Bruno, Improving diagnosis: Advances in radiology, Diagnosis 12 (2025) 578–587.

[2] F. Pesapane, G. Gnocchi, C. Quarrella, A. Sorce, L. Nicosia, L. Mariano, A. C. Bozzini, I. Marinucci,

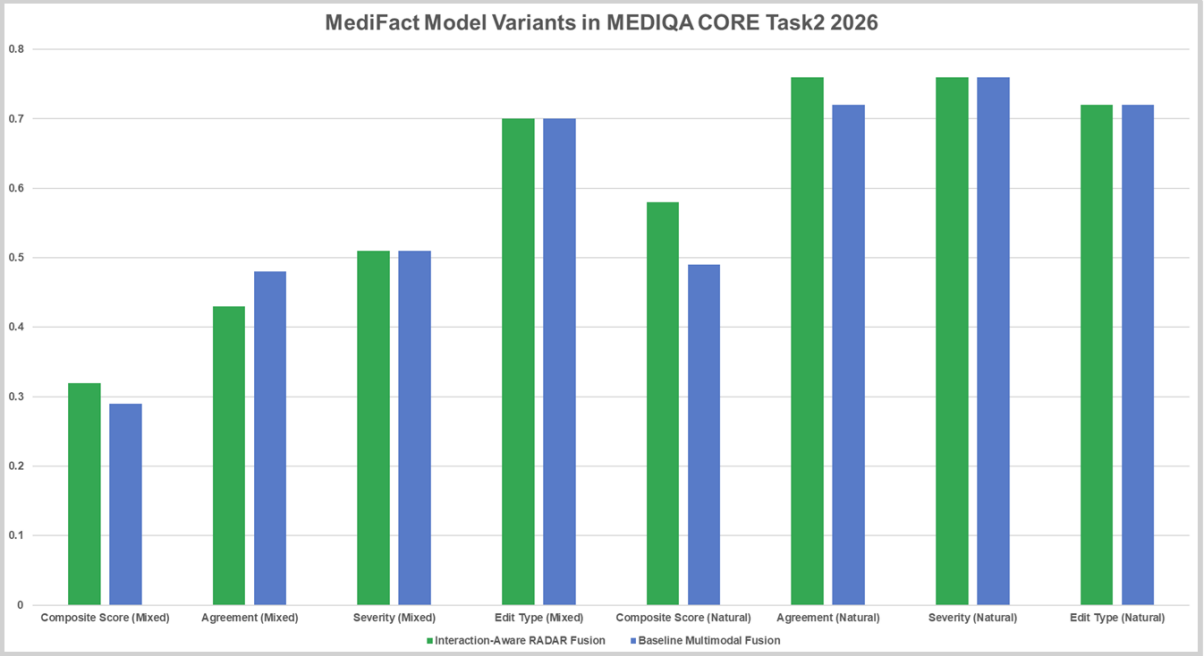

Figure 4: Comparison between the baseline multimodal fusion approach and the proposed interaction-aware RADAR fusion framework. The proposed approach demonstrates improved overall performance, particularly in the Natural evaluation setting, highlighting the effectiveness of interaction-aware feature fusion and balanced multimodal learning.

F. Priolo, F. Abbate, et al., Errors in radiology: A standard review, Journal of Clinical Medicine 13 (2024) 4306.

[3] B. D. Simon, K. B. Ozyoruk, D. G. Gelikman, S. A. Harmon, B. Türkbey, The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review, Diagnostic and Interventional Radiology 31 (2025) 303.

[4] B. Jandoubi, M. A. Akhloufi, Multimodal artificial intelligence in medical diagnostics, Information 16 (2025) 591.

[5] S. A. Merchant, N. Merchant, S. L. Varghese, M. J. S. Shaikh, Large language models and large concept models in radiology: Present challenges, future directions, and critical perspectives, World Journal of Radiology 17 (2025) 114754.

[6] Z. Sun, M. Jagtiani, W.-w. Yim, F. Xia, M. Gunn, M. Yetisgen, A. B. Abacha, Radar: A multimodal benchmark for 3d image-based radiology report review, arXiv preprint arXiv:2603.06681 (2026).

[7] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al., Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, arXiv preprint arXiv:2507.06261 (2025).

[8] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).

[9] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryılmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of imageclef 2026: Multimodal challenges in medicine, science, agritech, and security, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science

LNCS, Jena, Germany, 2026.
[10] A. Ben Abacha, Z. Sun, W. Yim, F. Xia, M. Yetisgen, Overview of the mediqa-core 2026 task 2 on radiology report discrepancy assessment, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[11] H. Xuyan, S. Meng, S. Chengxing, L. Haoxuan, Z. Jianlin, Visual-language reasoning large language models for primary care: advancing clinical decision support through multimodal ai: X. huang et al., The Visual Computer 41 (2025) 11327–11348.
[12] H. Ahmadi, E. Santamaría-Vázquez, L. Mesin, R. Hornero, Semantic-aware decoding of covert inner speech: A multimodal eeg–emg–audio framework (2026).
[13] X. Xu, J. Li, Z. Zhu, L. Zhao, H. Wang, C. Song, Y. Chen, Q. Zhao, J. Yang, Y. Pei, A comprehensive review on synergy of multi-modal data and ai technologies in medical diagnosis, Bioengineering 11 (2024) 219.
[14] P. Sloan, P. Clatworthy, E. Simpson, M. Mirmehdi, Automated radiology report generation: A review of recent advances, IEEE Reviews in Biomedical Engineering 18 (2024) 368–387.
[15] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).
[16] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, IEEE transactions on pattern analysis and machine intelligence 35 (2012) 221–231.
[17] X. Zeng, A. Ahmed, M. H. Tunio, A balanced multimodal multi-task deep learning framework for robust patient-specific quality assurance, Diagnostics 15 (2025) 2555.
[18] D. Mason, Su-e-t-33: pydicom: an open source dicom library, Medical Physics 38 (2011) 3493–3493.

Online Resources

The implementation code for the proposed MediFact framework is publicly available at:

- GitHub Repository: https://github.com/NadiaSaeed/Medifact-MEDIQA-CORE-Task-2-2026
- ImageCLEFmed MEDIQA 2026 Task Page: https://ai4media-bench.aimultimedialab.ro/competitions/7/
- Official Evaluation Script: https://github.com/GeraldSun/RADAR