IReL_IIT(BHU) at MEDIQA-CORE-Task-2 2026: Hybrid Multimodal Learning for Radiology Report Discrepancy Analysis

Krishna Tewari[1,*,†], Suhani Verma[2,†] and Sukomal Pal[1]

[1]Indian Institute of Technology (BHU) Varanasi, India
[2]Banasthali Vidyapith, Rajasthan, India

Abstract

Radiology report discrepancy analysis is an important task for improving diagnostic reliability, reducing reporting inconsistencies, and supporting clinical decision-making workflows. In this work, we present IReL_IIT(BHU)'s submission to the MEDIQA-CORE Task 2 at CLEF 2026, focusing on automated agreement classification between radiology reports and suggested edits. We propose a hybrid multimodal framework integrating transformer-based biomedical language modeling, multimodal feature fusion, hierarchical reasoning, synthetic data augmentation, and rule-based clinical inference. The proposed system utilizes PubMedBERT for contextual textual representation learning, while multimodal variants additionally incorporate CT scan representations extracted using 3D convolutional neural networks. Reliability-aware hierarchical prediction is further introduced for confidence-guided agreement classification, while synthetic radiology edit generation and domain-specific contradiction reasoning are employed to improve robustness and semantic consistency. The systems are evaluated on the MEDIQA-RADAR benchmark using agreement classification metrics, confusion matrices, and official leaderboard evaluation protocols. Experimental results demonstrate that the proposed framework achieves competitive performance for radiology discrepancy analysis, obtaining a development F1 of 71.82%. On the official leaderboard, the system achieved a Composite Score of 0.05 on the Mixed Set and 0.08 on the Natural Set, while demonstrating comparatively stronger severity prediction performance with scores of 0.51 and 0.76 respectively. Despite these improvements, the framework remains limited by class imbalance, dependence on synthetic rule generation, and reduced generalization across highly diverse clinical edits. The source code for all experiments will be made publicly available at https://github.com/cse-iitbhu/Report-Discrepancy-Summarization.

Keywords

Radiology Report Discrepancy Analysis, Medical NLP, Multimodal Learning, Synthetic Data Augmentation

1. Introduction and Related Work

Radiology reports are a critical component of clinical decision-making, providing structured interpretations of medical imaging examinations. It guides diagnosis, treatment planning, and patient management. The increasing volume and complexity of medical imaging studies have motivated the development of artificial intelligence (AI) systems capable of assisting radiologists in image interpretation, report generation, and quality assurance. Recent advances in deep learning, vision-language learning, and large language models have significantly improved the ability of AI systems to jointly process medical images and clinical text, enabling a wide range of applications including report generation, visual question answering, image-text retrieval, and clinical decision support [1, 2].

The emergence of large-scale medical datasets and multimodal pretraining strategies has accelerated research in automated radiology reporting. Transformer-based architectures have established strong baselines for medical report generation by leveraging cross-modal attention mechanisms between visual features and textual representations [3, 4].

Recent years have also witnessed the rapid emergence of medical foundation models capable of

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
krishnatewari.rs.cse24@iitbhu.ac.in (K. Tewari); suhani.fgc@gmail.com (S. Verma); spal.cse@iitbhu.ac.in (S. Pal)
0009-0005-6599-9956 (K. Tewari); 0000-0001-8743-9830 (S. Pal)

performing diverse clinical reasoning tasks. Models such as BioMedGPT [5], Med-Flamingo [6], LLaVA-Med [7], RadFM [8], Med-PALM M [9], and PMC-LLaMA [10] have demonstrated strong performance across medical image understanding, report generation, visual question answering, and multimodal clinical reasoning tasks. These models leverage large-scale pretraining and instruction tuning to integrate visual and textual information, enabling increasingly sophisticated reasoning capabilities across a broad spectrum of biomedical applications.

Challenge-driven research has further demonstrated the effectiveness of multimodal foundation models in specialized clinical domains. Recent studies have explored CLIPSeg-based architectures for dermatological image understanding and visual question answering [11], while Florence-2 and diffusion-based frameworks have shown promising results for gastrointestinal image analysis, question answering, and synthetic image generation [12]. Additionally, confidence-aware multimodal reasoning approaches have been investigated for explainable visual question answering in clinical diagnostic settings, emphasizing the importance of transparency and reliability in medical AI systems [13].

Despite these advances, ensuring the correctness and clinical reliability of generated or revised radiology reports remains a significant challenge. Existing report generation systems frequently suffer from factual inconsistencies, hallucinated findings, omitted abnormalities, and clinically inaccurate statements, limiting their applicability in safety-critical healthcare environments [1, 2]. Consequently, increasing attention has been directed toward factual verification and clinically informed evaluation methodologies. Several studies have proposed specialized frameworks such as RadGraph [14], RadGraph-XL [15], RadCliQ [16]. While these approaches provide valuable tools, they primarily focus on complete report assessment and do not explicitly address report revision and discrepancy analysis.

In clinical practice, radiology reporting commonly follows a hierarchical review process in which preliminary reports prepared by trainees are reviewed and revised by experienced radiologists. These revisions may involve correcting inaccurate findings, incorporating omitted observations, or clarifying ambiguous descriptions. Assessing the validity and clinical significance of such modifications requires joint reasoning over imaging evidence, report context, and the semantic implications of proposed edits. Unlike traditional report generation tasks, discrepancy assessment demands fine-grained multimodal reasoning, requiring models to determine whether a candidate edit is supported by image evidence, estimate its potential clinical impact, and categorize the nature of the revision.

The application of recent methods to radiology report discrepancy assessment remains largely unexplored, particularly in settings involving volumetric abdominal computed tomography (CT) examinations and clinically meaningful report edits. Such scenarios require not only accurate image understanding but also nuanced clinical reasoning regarding the consequences of reporting errors and omissions.

The ImageCLEFmed MEDIQA-CORE 2026 Task 2 as part of ImageCLEF [17] benchmark, addresses this emerging research problem by focusing on multimodal radiology report review and discrepancy evaluation. Participants are required to assess candidate report edits using abdominal CT examinations and preliminary radiology reports, determining whether proposed modifications are supported by imaging findings, evaluating the severity of discrepancies, and classifying the type of revision. By emphasizing report verification rather than report generation alone, the task promotes the development of trustworthy multimodal systems capable of functioning as intelligent reviewers within radiology reporting workflows. The challenge therefore provides an important benchmark for advancing clinically grounded multimodal reasoning and quality assurance in medical imaging.

The rest of the paper is structured as follows: Section 2 describes the task overview and dataset employed; Section 3 presents the different proposed methodologies for the subtask; Section 4 reports results; and Section 5 concludes with key findings.

2. Task Overview and Dataset

The ImageCLEFmed MEDIQA-CORE 2026 Task 2: Report Discrepancy Assessment challenge focuses on multimodal radiology report review and discrepancy analysis [18]. The task is designed to emulate real-

world clinical reporting workflows in which preliminary reports generated by junior radiologists are subsequently reviewed and revised by experienced attending radiologists. Rather than generating reports from scratch, participating systems are required to evaluate proposed modifications and determine their validity and clinical importance.

Each sample consists of three components: (i) a three-dimensional abdominal computed tomography (CT) examination, (ii) a preliminary radiology report describing the examination, and (iii) a candidate report edit representing a proposed modification to the original report. The candidate edit may introduce new findings, correct existing statements, or clarify previously reported observations.

Given these inputs, the challenge requires models to perform three complementary prediction tasks. First, systems must determine the image-level agreement between the proposed edit and the underlying CT examination, assessing whether the modification is supported by visual evidence. Second, models must estimate the clinical severity of the discrepancy, reflecting the potential impact of the modification on diagnostic interpretation and patient management. Third, systems must classify the discrepancy type by identifying whether the proposed modification corresponds to a correction, an addition, or a clarification.

2.1. Dataset

The dataset [19] provided for the ImageCLEFmed MEDIQA-CORE 2026 Report Discrepancy Assessment task consists of abdominal CT examinations, preliminary radiology reports, clinical metadata, and candidate report edits designed to simulate real-world radiology review workflows. Each case is uniquely identified by a 'case_id' (e.g., 10001) and includes a clinical indication describing the reason for imaging, metadata such as the preliminary reporting time, and a set of candidate edits stored in 'test_edits.json'. Each candidate edit is associated with a unique 'edit_id' and a corresponding 'suggested_edit_text' describing a proposed modification to the preliminary report. For example, Case 10001 contains the clinical indication

- "small bowel mass seen on push, eval for neoplasm, bleeding"

and includes multiple candidate edits such as

- "There is evidence of small bowel obstruction with proximal dilatation",
- "Bulky mesenteric lymphadenopathy is present adjacent to the jejunal mass", and
- "The gallbladder is distended with multiple gallstones and wall thickening"

The corresponding preliminary report is intentionally concise, stating only that there is

- "No evidence of active intestinal hemorrhage on this limited CT" and
- "Mass seen on prior endoscopy is not well seen on this CT"

This substantial semantic gap between the original report and the proposed edits reflects authentic report revision scenarios in which attending radiologists may add omitted findings, correct inaccuracies, or provide more detailed interpretations. The imaging component of the dataset is organized into case-specific directories containing multiple DICOM series rather than a single CT volume. For instance, the directory for Case 10001 contains reconstruction folders such as 'C-A-P,_W_IV', 'CHEST_MIP', 'COR_BODY', 'LUNG_ALG', 'SAG_BODY', and 'SCOUT', each representing different acquisition protocols, reconstruction kernels, or anatomical views. These multi-series studies require models to reason across heterogeneous image representations and associate imaging evidence with textual report modifications. Consequently, the task is formulated as a multimodal discrepancy assessment problem in which each candidate edit is evaluated independently using the CT examination, preliminary report, and clinical context to determine whether the proposed modification is supported by the imaging findings, assess its clinical significance, and classify its discrepancy type.

3. Methodology

The proposed work presents three progressively improved hybrid radiology agreement classification pipelines combining transformer-based biomedical language modeling, multimodal learning, synthetic data augmentation, hierarchical reasoning, and rule-based clinical reasoning.

3.1. Approach 1

Figure 1 illustrates the overall architecture of our multimodal discrepancy assessment framework. Each candidate edit is treated as an independent training instance constructed from the clinical indication, preliminary radiology report, and suggested report modification. For every edit, the corresponding CT examination is loaded from the available DICOM series and combined with textual information to form a multimodal representation. The framework jointly processes volumetric CT data and clinical text using dedicated image and language encoders, followed by a shared fusion module and multiple task-specific classification heads. The model simultaneously predicts agreement, severity, and edit type through a multi-task learning strategy.

3.1.1. Data Preparation

For each case, the clinical indication, preliminary report, and candidate edit text are concatenated into a single textual sequence using the format:

```
Clinical Indication [SEP] Preliminary Report [SEP] Suggested Edit
```

Each candidate edit is considered an independent sample. The textual input is tokenized using the PubMedBERT tokenizer and truncated or padded to a maximum sequence length of 256 tokens.

To generate supervision signals for training, rule-based labeling functions are employed. Agreement labels are derived using lexical overlap between the preliminary report and the candidate edit. The overlap ratio determines whether an edit is assigned to the categories disagree, partially agree, or agree. Edit type labels are generated using keyword matching, where terms such as “no”, “absent”, and “not seen” are mapped to correction, terms such as “add”, “also”, and “additional” are mapped to addition, and all remaining edits are assigned to clarification. Severity labels are similarly obtained through keyword matching, where edits mentioning terms such as “mass”, “tumor”, “bleed”, or “hemorrhage” are assigned critical severity, findings such as “lesion”, “infection”, or “inflammation” are assigned moderate severity, and all other edits are categorized as negligible.

These heuristic labeling functions were designed to provide weak supervision in the absence of manually annotated auxiliary labels. Since the generated labels rely on lexical overlap and predefined keyword rules rather than image-grounded clinical reasoning, they may introduce noisy supervision for complex edits involving implicit findings, paraphrases, or nuanced clinical interpretations.

3.1.2. CT Volume Processing

The imaging component is constructed directly from the DICOM studies associated with each case. For every examination, the framework selects the first available image series within the corresponding case directory. Individual DICOM slices are read using the pydicom library, and only images with CT modality are retained. If a slice contains multiple channels, only a single channel is preserved.

The first 50 valid slices are stacked to form a volumetric representation. A fixed number of slices was selected to standardize the volumetric input across examinations while keeping GPU memory usage and computational cost manageable. This preprocessing strategy enabled efficient batch training with a consistent input size for the 3D convolutional encoder. Intensity values are normalized using min-max normalization and subsequently resized using trilinear interpolation to a fixed resolution of $64 \times 128 \times 128$ voxels. When no valid CT slices are available, a zero-valued volume of identical dimensions is used as a fallback representation.

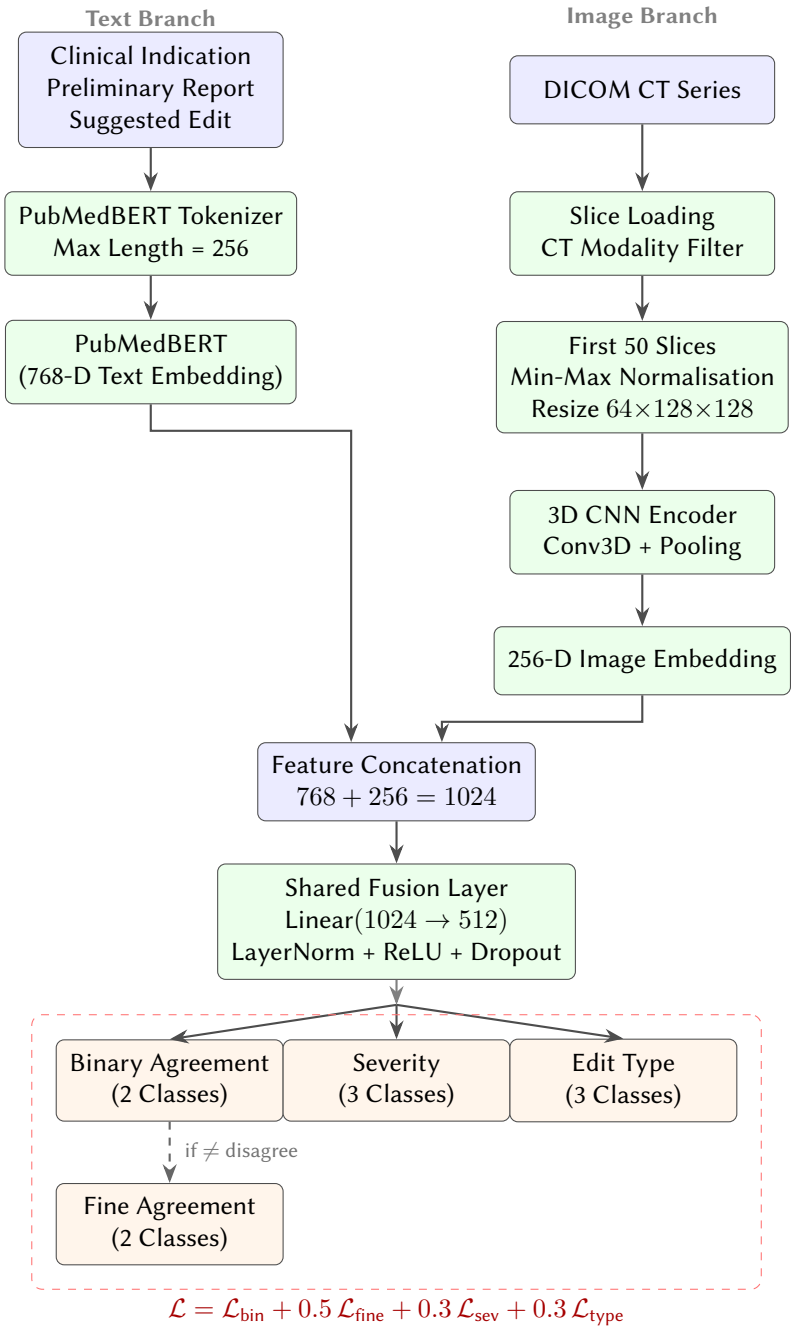

Figure 1: Overview of Approach 1

3.1.3. Multimodal Architecture

The textual branch utilizes the PubMedBERT [1] model as the language encoder. The contextual representation corresponding to the first token is extracted from the final hidden layer, resulting in a 768-dimensional text embedding.

For image processing, a lightweight three-dimensional convolutional neural network is employed. The network consists of two consecutive 3D convolutional layers with ReLU activations and max-pooling operations, followed by adaptive average pooling and a fully connected projection layer. The final image representation is projected to a 256-dimensional feature vector.

The text and image embeddings are concatenated and passed through a shared fusion module

[1] `microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext`

consisting of a fully connected layer, layer normalization, ReLU activation, and dropout. The resulting multimodal representation is used as input to four task-specific prediction heads.

3.1.4. Multi-Task Learning

The model jointly optimizes three classification objectives:

- Agreement classification.
- Severity prediction.
- Edit type prediction.

The binary agreement classifier predicts whether a candidate edit disagrees with the preliminary report or exhibits some level of agreement. The fine-grained classifier further distinguishes partially agree and agree categories. Separate classification heads are used for severity assessment and edit type prediction.

During training, the overall loss is computed as a weighted combination of cross-entropy losses i.e., $\mathcal{L} = \mathcal{L}_{bin} + 0.5\mathcal{L}_{fine} + 0.3\mathcal{L}_{sev} + 0.3\mathcal{L}_{type}$. This formulation encourages the model to learn a shared multimodal representation while simultaneously optimizing all discrepancy assessment objectives. The coefficients used to combine the individual task losses were selected heuristically to prioritize the primary agreement prediction task while allowing the severity and edit-type objectives to act as auxiliary supervision. No extensive hyperparameter tuning of these coefficients was performed on the development set.

To mitigate the severe class imbalance present in the agreement labels, class-weighted cross-entropy was employed during optimization. The class weights were computed automatically from the training-set class frequencies using inverse-frequency weighting and were not manually tuned.

During prediction, the binary agreement classifier is evaluated first. If the predicted class corresponds to disagreement, the final agreement label is assigned directly. Otherwise, the output of the fine-grained classifier is used to distinguish between partially agree and agree categories. Severity and edit type predictions are obtained from their respective classification heads. The final system generates three outputs for each candidate edit: agreement prediction, severity prediction, and edit type prediction.

3.2. Approach 2

Approach 2 builds upon the multimodal architecture described in Approach 1 and retains the same text encoder, CT processing pipeline, feature fusion strategy, and multi-task learning framework. The primary modification is introduced during agreement prediction, where threshold-aware hierarchical inference is employed.

3.2.1. Refined Rule-Based Label Generation

Compared with Approach 1, a simplified set of heuristic rules is used for generating auxiliary supervision labels. Edit type labels are assigned using only two keywords. Edits containing the term “no" are mapped to the correction category, edits containing the term “add" are assigned to the addition category, and all remaining edits are categorized as clarification.

Similarly, severity labels are generated using a reduced keyword vocabulary. Edits containing “tumor" or “bleed" are assigned critical severity, while edits mentioning “lesion" or “infection" are categorized as moderate severity. All other edits are labelled as negligible.

3.2.2. Modified Fusion Layer

The multimodal fusion module is simplified by removing the layer normalization component used in Approach 1. The concatenated text and image embeddings are projected into a shared 512-dimensional representation through a fully connected layer followed by ReLU activation and dropout: $1024 \rightarrow 512$. This shared representation is subsequently used by the agreement, severity, and edit-type prediction heads.

3.2.3. Threshold-Based Hierarchical Inference

The most significant modification in Approach 2 is the introduction of threshold-aware agreement prediction. Rather than directly using the binary agreement classifier output, the softmax probability assigned to the disagreement class is explicitly considered.

Let $P_{disagree}$ denote the probability assigned to the disagreement category by the binary agreement classifier. If:

$$P_{disagree} > 0.7,$$

the final agreement label is directly assigned as *disagree*. Otherwise, the sample is forwarded to the fine-grained agreement classifier, which distinguishes between *partially agree* and *agree*. The category with the larger softmax probability is selected as the final prediction. This threshold-based decision mechanism allows the model to identify high-certainty disagreement cases before performing finer agreement discrimination, thereby reducing error propagation between the two stages of the hierarchical framework.

The threshold value of 0.7 was selected empirically as a conservative confidence cutoff for identifying high-certainty disagreement predictions. The objective was to reduce error propagation from the binary classifier to the fine-grained agreement classifier by forwarding only sufficiently confident disagreement cases. A systematic comparison of alternative threshold values was not performed and remains an important direction for future work.

3.3. Approach 3

Unlike Approaches 1 and 2, Approach 3 performs agreement prediction using only textual information. CT images are not utilized during training or inference. The framework as shown in Figure 2 combines synthetic training sample generation, class-weighted optimization, and rule-based prediction refinement.

3.3.1. Synthetic Data Generation

To increase the number of training samples, synthetic edit instances are generated for every radiology report. A predefined list of findings (mass, lesion, tumor, fracture, nodule and opacity) is used to create additional candidate edits. For each finding, six synthetic edits are generated:

- There is a finding
- Finding is present
- No finding seen
- No evidence of finding
- Possible small finding
- Mild finding suspected

The agreement labels for the synthetic edits are assigned deterministically using the predefined generation templates rather than manual annotation. For each predefined radiological finding (e.g., mass, lesion, tumor, fracture, nodule, and opacity), templates expressing positive findings (e.g., "There is a mass" and "Mass is present") are labeled as disagree, explicit negation templates (e.g., "No mass seen" and "No evidence of mass") are labeled as agree, while uncertain expressions (e.g., "Possible small mass" and "Mild mass suspected") are labeled as partially agree. These synthetic examples are intended only as auxiliary training samples to improve robustness under limited training data.

3.3.2. Text Classification Model

The textual encoder is same as Approach 1 and 2. For each sample, the clinical indication, preliminary report, and suggested edit are concatenated into a single sequence and tokenized using a maximum sequence length of 256 tokens. The contextual representation corresponding to the first token is extracted from the final transformer layer and passed through a lightweight classification network

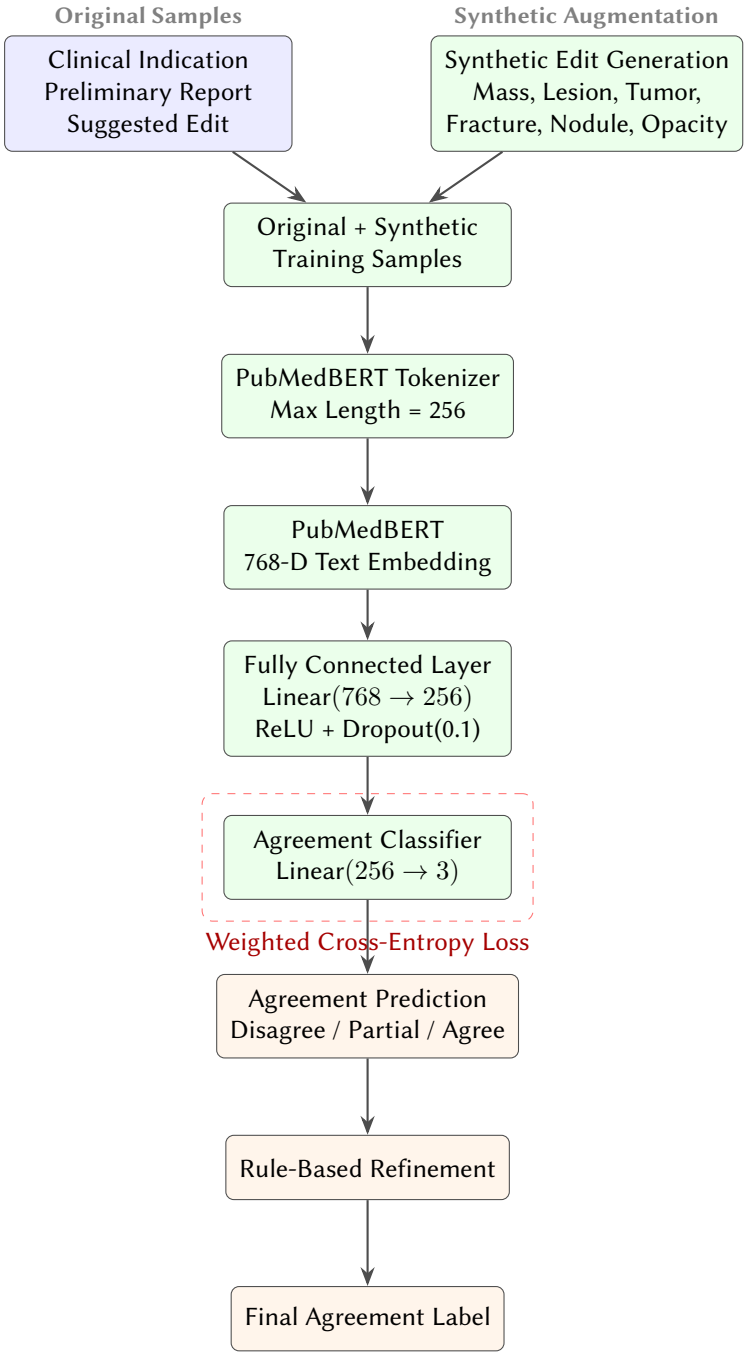

Figure 2: Overview of Approach 3

consisting of $768 \to 256 \to 3$, where the intermediate layer contains a fully connected projection, ReLU activation, and dropout with a probability of 0.1. The final layer directly predicts the three agreement categories: disagree, partially agree, and agree.

During training, only parameters belonging to the final transformer layer of PubMedBERT are updated. All other transformer parameters remain frozen.

Class frequencies are computed from the training set and inverse-frequency weights are assigned to each agreement class. These weights are incorporated into the cross-entropy loss function to reduce the effect of class imbalance during optimization.

3.3.3. Rule-Based Prediction Refinement

After the neural classifier produces a prediction, a set of predefined rules is evaluated using the report and candidate edit.

The disagreement category is assigned when:

- The report contains “no evidence of" and the edit contains “present", “seen", or “identified";
- The report contains “normal" and the edit contains “mass", “lesion", “tumor", or “abnormal";
- The report contains “unremarkable" and the edit contains “mass", “lesion", or “abnormal".

The agreement category is assigned when the edit contains any of: new/increase/worsening/progression/consistent with/suggesting. The partially agree category is assigned when the edit contains any of: mild/small/slight/possible. If none of the rules are triggered, the final prediction is obtained directly from the neural classifier.

3.4. Training Details

The proposed multimodal framework was implemented using PyTorch. The text encoder was initialized with PubMedBERT, while the image branch employed the lightweight 3D CNN described in Section 3.1.3. The model was optimized using the AdamW optimizer with a learning rate of 1×10^{-5}. Training was performed for 8 epochs with a batch size of 4. A maximum input sequence length of 256 tokens was used for the textual encoder. To mitigate the effects of class imbalance, class-weighted cross-entropy loss was employed, where the class weights were computed automatically from the inverse class frequencies of the training set. The dataset was randomly partitioned into training and validation subsets using an 80:20 split for model development and performance monitoring.

4. Results and Discussions

This section presents the results obtained in the subtasks across different evaluation metrics mentioned in the overview paper of the task [18].

Table 1 presents the performance of our submitted approaches on the MEDIQA-RADAR benchmark. Among our submissions, Approach 1 achieved the best overall performance with a Mixed Composite Score of 0.05 and a Natural Composite Score of 0.08. Approaches 2 and 3 obtained lower composite scores, achieving Mixed Composite Scores of 0.03 and 0.03 respectively.

A detailed comparison of the individual metrics reveals that severity prediction was the strongest component across all submissions. Both Approaches 1 and 3 achieved a Mixed Severity Score of 0.51, while Approaches 1 and 2 obtained Natural Severity Scores of 0.76. This suggests that the severity labeling strategy captured relatively consistent patterns that generalized reasonably well across evaluation subsets.

Agreement prediction exhibited considerably greater variability. Approach 3 achieved the highest Mixed Agreement Score of 0.46, while Approach 2 obtained 0.26 and Approach 1 achieved 0.22. However, this improvement did not translate into a higher composite score because the gains in agreement prediction were accompanied by substantial reductions in edit-type performance. On the Natural subset, agreement performance remained low across all approaches, indicating that accurately assessing report-edit agreement remains a challenging problem.

The strongest edit-type performance was obtained by Approach 2, achieving Mixed and Natural Edit Type Scores of 0.38 and 0.39 respectively. In comparison, Approaches 1 and 3 obtained considerably lower edit-type scores. This observation suggests that the threshold-based hierarchical strategy introduced in Approach 2 was more effective for identifying edit categories than the alternative architectures explored in the other submissions.

Table 1
Leaderboard entries visible in the official evaluation results. Scores are reported for both the Mixed and Natural subsets. The best value in each column is highlighted in bold.

Team	Run ID	Mixed Set				Natural Set			
		Comp.	Agr.	Sev.	Type	Comp.	Agr.	Sev.	Type
MEDIQA Organizers	797	0.39	0.68	0.56	0.82	0.48	0.63	0.61	0.84
anubhav27	1031	0.33	0.52	0.52	0.68	0.55	0.82	0.72	0.71
anubhav27	1028	0.32	0.48	0.51	0.70	0.58	0.86	0.76	0.72
anubhav27	1030	0.24	0.52	0.41	0.61	0.34	0.52	0.52	0.66
anubhav27	1027	0.20	0.54	0.34	0.58	0.22	0.47	0.33	0.63
nadiasaeed	1043	0.32	0.43	0.51	0.70	0.58	0.76	0.76	0.72
nadiasaeed	1018	0.29	0.48	0.51	0.70	0.49	0.72	0.76	0.72
nadiasaeed	967	0.21	0.55	0.52	0.70	0.31	0.51	0.76	0.71
nadiasaeed	1020	0.20	0.53	0.41	0.70	0.27	0.47	0.53	0.72
nadiasaeed	1045	0.06	0.30	0.24	0.70	0.00	0.26	0.07	0.72
nadiasaeed	1033	0.02	0.40	0.34	0.22	0.01	0.07	0.48	0.14
ClinicalVLM-TW	1069	0.22	0.41	0.38	0.64	0.25	0.22	0.42	0.69
ClinicalVLM-TW	1037	0.22	0.53	0.38	0.63	0.29	0.49	0.46	0.65
ClinicalVLM-TW	1034	0.21	0.53	0.38	0.60	0.27	0.49	0.46	0.64
ClinicalVLM-TW	1092	0.21	0.51	0.36	0.63	0.25	0.41	0.42	0.67
ClinicalVLM-TW	1013	0.19	0.42	0.34	0.63	0.22	0.37	0.34	0.65
ClinicalVLM-TW	942	0.17	0.51	0.34	0.61	0.24	0.56	0.40	0.66
ClinicalVLM-TW	836	0.17	0.51	0.34	0.61	0.24	0.56	0.40	0.66
ClinicalVLM-TW	1068	0.14	0.49	0.32	0.62	0.16	0.27	0.39	0.67
IReL_IIT(BHU)	**960**	**0.05**	**0.22**	**0.51**	**0.12**	**0.08**	**0.24**	**0.76**	**0.14**
IReL_IIT(BHU)	**957**	**0.03**	**0.26**	**0.29**	**0.38**	**0.06**	**0.16**	**0.30**	**0.39**
IReL_IIT(BHU)	**895**	**0.03**	**0.46**	**0.51**	**0.11**	**0.03**	**0.03**	**0.76**	**0.18**

4.1. Limitations

The most significant challenge encountered during experimentation was the severe class imbalance present in the dataset. The agreement categories were highly skewed, causing the models to favor majority classes during training. Although several mitigation strategies were explored, including hierarchical classification, threshold-based inference, class-weighted optimization, and synthetic data augmentation, the imbalance remained a major obstacle and substantially affected agreement prediction performance.

For the multimodal approaches, only the first available CT series and the first 50 valid slices were processed. While this design reduced computational cost and enabled a fixed-size volumetric representation, diagnostically important findings may appear in later slices or alternative reconstruction series that were not considered. Consequently, some clinically relevant information may have been excluded, potentially limiting agreement prediction performance. Future work will investigate automatic series selection, adaptive slice sampling, and full-volume vision encoders to better exploit the complete CT examination.

The synthetic examples introduced in Approach 3 were generated using a small set of manually designed templates. While this increased the number of available training samples, the generated edits did not adequately reflect the linguistic diversity and complexity of real radiology report modifications, limiting the effectiveness of the augmentation strategy.

Moreover, the auxiliary agreement, severity, and edit-type labels were generated using simple rule-based heuristics. Although these labels provide additional supervision during training, they do

not fully capture the semantic complexity of radiology report revisions or image-grounded clinical reasoning. Consequently, label noise may have affected representation learning and reduced the overall performance of the multi-task model.

5. Conclusion and Future Work

In this work, we investigated multiple approaches for the MEDIQA-RADAR task of radiology discrepancy assessment. Three different systems were developed and evaluated, including multimodal architectures that combine radiology images and textual reports, confidence-based hierarchical agreement prediction, and a text-only framework incorporating synthetic data augmentation and rule-based refinement. Experimental results demonstrated that the multimodal approach achieved the best overall performance among our submissions, highlighting the potential benefit of leveraging both visual and textual information for discrepancy assessment. However, agreement prediction remained substantially more challenging than severity and edit-type classification, indicating the complexity of identifying semantic consistency between candidate edits and radiology reports.

Several challenges limited overall performance. Most notably, severe class imbalance in the agreement categories negatively affected model learning despite the use of class weighting, hierarchical inference, and synthetic augmentation. In addition, the automatically generated supervision signals and lightweight image encoders may have restricted the quality of the learned representations.

In future work, we plan to investigate stronger multimodal architectures based on vision-language foundation models and more advanced volumetric image encoders. Exploring contrastive multimodal pretraining, improved fusion strategies, and clinically informed prompting techniques may further enhance performance. We also aim to develop more realistic synthetic data generation methods and employ expert-annotated supervision to reduce label noise and improve generalization across diverse radiology discrepancy scenarios.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] X. Wang, G. Figueredo, R. Li, W. E. Zhang, W. Chen, X. Chen, A survey of deep-learning-based radiology report generation using multimodal inputs, Medical Image Analysis 103 (2025) 103627. doi:10.1016/j.media.2025.103627.

[2] A. Izhar, N. Idris, N. Japar, Medical radiology report generation: A systematic review of current deep learning methods, trends, and future directions, Artificial Intelligence in Medicine 168 (2025) 103220. doi:10.1016/j.artmed.2025.103220.

[3] Z. Chen, Y. Shen, Y. Song, X. Wan, Cross-modal memory networks for radiology report generation, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 5904–5914. URL: https://aclanthology.org/2021.acl-long.459/. doi:10.18653/v1/2021.acl-long.459.

[4] J.-B. Delbrouck, P. Chambon, C. Bluethgen, E. Tsai, O. Almusa, C. Langlotz, Improving the factual correctness of radiology report generation with semantic rewards, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 4348–4360. URL:

https://aclanthology.org/2022.findings-emnlp.319/. doi:10.18653/v1/2022.findings-emnlp.319.

[5] Y. Luo, J. Zhang, S. Fan, K. Yang, M. Hong, Y. Wu, M. Qiao, Z. Nie, BioMedGPT: An open multimodal large language model for BioMedicine, IEEE J. Biomed. Health Inform. 30 (2026) 981–992.

[6] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, P. Rajpurkar, Med-flamingo: a multimodal medical few-shot learner, in: S. Hegselmann, A. Parziale, D. Shanmugam, S. Tang, M. N. Asiedu, S. Chang, T. Hartvigsen, H. Singh (Eds.), Proceedings of the 3rd Machine Learning for Health Symposium, volume 225 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 353–367. URL: https://proceedings.mlr.press/v225/moor23a.html.

[7] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, J. Gao, Llava-med: training a large language-and-vision assistant for biomedicine in one day, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[8] C. Wu, X. Zhang, Y. Zhang, H. Hui, Y. Wang, W. Xie, Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data, Nat. Commun. 16 (2025) 7866.

[9] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, A. Palepu, B. Mustafa, A. Chowdhery, Y. Liu, S. Kornblith, D. Fleet, P. Mansfield, S. Prakash, R. Wong, S. Virmani, C. Semturs, S. S. Mahdavi, B. Green, E. Dominowska, B. A. y Arcas, J. Barral, D. Webster, G. S. Corrado, Y. Matias, K. Singhal, P. Florence, A. Karthikesalingam, V. Natarajan, Towards generalist biomedical ai, NEJM AI 1 (2024) AIoa2300138. URL: https://ai.nejm.org/doi/full/10.1056/AIoa2300138. doi:10.1056/AIoa2300138. arXiv:https://ai.nejm.org/doi/pdf/10.1056/AIoa2300138.

[10] C. Wu, W. Lin, X. Zhang, Y. Zhang, Y. Wang, W. Xie, Pmc-llama: Towards building open-source language models for medicine, arXiv preprint arXiv:2304.14454 (2023). URL: https://arxiv.org/abs/2304.14454. doi:10.48550/arXiv.2304.14454.

[11] K. Tewari, A. Verma, S. Pal, IReL, IIT(BHU) at MEDIQA-MAGIC 2025: Tackling Multimodal Dermatology with CLIPSeg-Based Segmentation and BERT-Swin Question Answering, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS, 2025. URL: https://ceur-ws.org/Vol-4038/paper_207.pdf.

[12] K. Tewari, S. Pal, Advancing Vision and Language in GI Diagnosis: Florence2 for Question Answering and Stable Diffusion for Image Synthesis, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS, 2025. URL: https://ceur-ws.org/Vol-4038/paper_208.pdf.

[13] K. Tewari, S. Pal, X-VQA for GI Diagnostics: Multimodal Visual Question Answering with Confidence-Aware Explanations, in: Proceedings of the MediaEval 2025 Multimedia Benchmark Workshop, 2025. URL: https://2025.multimediaeval.com/paper38.pdf.

[14] S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. Nguyen Duong, T. Bui, P. Chambon, M. Lungren, A. Ng, C. Langlotz, P. Rajpurkar, RadGraph: Extracting Clinical Entities and Relations from Radiology Reports, PhysioNet (2021). URL: https://doi.org/10.13026/hm87-5p47. doi:10.13026/hm87-5p47, version 1.0.0.

[15] J.-B. Delbrouck, RadGraph-XL: A Large-Scale Expert-Annotated Dataset for Entity and Relation Extraction from Radiology Reports, PhysioNet (2025). URL: https://doi.org/10.13026/j8e7-pr22. doi:10.13026/j8e7-pr22, version 1.0.0.

[16] F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E. P. Reis, E. K. U. N. Fonseca, H. M. H. Lee, Z. S. H. Abad, A. Y. Ng, C. P. Langlotz, V. K. Venugopal, P. Rajpurkar, Evaluating progress in automatic chest x-ray radiology report generation, Patterns (N. Y.) 4 (2023) 100802.

[17] 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.

[18] 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.

[19] Z. Sun, M. Jagtiani, W. Yim, F. Xia, M. Gunn, M. Yetisgen, A. Ben Abacha, Radar: A multimodal benchmark for 3d image-based radiology report review, arXiv preprint arXiv:2603.06681 (2026). URL: https://arxiv.org/abs/2603.06681.