Learning from Textual Radiology Reports: A Benchmark Dataset for Coronary CT Angiography

Sudharshan Balaji[1] **Zhiyu Liu**[5*] **Zhengyuan Jiang**[1] **Shuo Lei**[2] **Yimin Chen**[3]
Yang Xiao[4] **Shone Almeida**[1] **Mathew Karivelil**[1] **Christopher Malanga**[1] **Ning Wang**[1]
[1]University of South Florida [2]Virginia Tech [3]University of Massachusetts, Lowell
[4]University of Kentucky [5]Baystate Medical Center
{sudharshanbalaji, ningw}@usf.edu zhiyu.liu@baystatehealth.org

Abstract

While coronary imaging is widely used for anatomical assessment, CCTA reports play a distinct last-mile role in clinical care. Rather than serving as an intermediate signal, CCTA provides an assessment of coronary disease severity (known as the CAD-RADS score) to guide patient management. However, real-world clinical text exhibits substantial heterogeneity in terminology and structure, leading to inconsistent interpretation by automated systems, even for clinically similar cases.

Recent work leverages a direct application of LLMs for automated CAD-RADS scoring, but is limited by small, non-public, and homogeneous clinical data. We introduce CCTA-RADS, the largest publicly available dataset of 940 real-world CCTA reports from Tampa General Hospital, each annotated with CAD-RADS scores. Our analysis reveals that direct approaches, including state-of-the-art LLMs (GPT-4o, GPT-o3) and fine-tuned BERT models underperform on diverse real-world clinical data. To address these limitations, we propose a two-stage pipeline that decouples structuring from classification: an LLM-based parser normalizes heterogeneous reports into structured JSON format, followed by fine-tuned BERT classification. This approach substantially improves the F1-score by 6%-13% compared with direct methods. We deploy our system as an interactive web interface that allows clinicians to upload CCTA reports for automated CAD-RADS assessment with SHAP and LIME explainability visualizations.

1 Introduction

Coronary CT Angiography (CCTA) plays a critical role in diagnosing Coronary Artery Disease (CAD), with the CAD-RADS scoring system providing standardized criteria for classifying stenosis severity and guiding patient management (Cury et al., 2016, 2022). As a decision-ready clinical artifact, CCTA reports translate anatomical findings into actionable risk stratification and management pathways. Therefore, automated CAD-RADS extraction from real-world CCTA reports has the potential to improve consistency, reduce inter-reader variability, and enhance the scalability and efficiency of clinical workflows.

However, real-world clinical CCTA reports present substantial challenges for automated analysis (Meystre et al., 2008; Ford et al., 2016). These documents exhibit significant heterogeneity in reporting styles, terminology variations, structural inconsistencies, and narrative formatting across institutions and individual radiologists (Cai et al., 2016). Reports may use different section headers, varied stenosis descriptions ("25% stenosis" vs. "mild narrowing"), inconsistent anatomical references, and diverse formatting conventions. This linguistic and structural variability makes it difficult for NLP models to reliably extract clinically relevant information for automated scoring.

Recent work has explored direct application of large language models for CAD-RADS classification (Wang et al., 2024; Arnold et al., 2025; Min et al., 2025). However, these approaches face critical limitations: (1) they rely on relatively small datasets (200-590 reports), making it difficult to assess how models generalize to data; (2) the datasets are non-public or synthetically augmented, which may not capture real-world complexity. We introduce CCTA-RADS, the largest publicly available dataset of 940 real-world CCTA reports from Tampa General Hospital, each expert-annotated

[*]Work done while the author was at USF.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 480–493

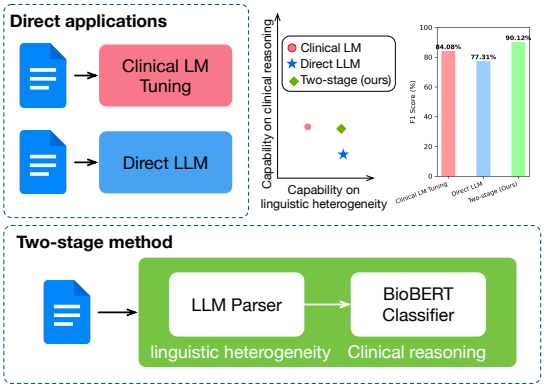

Figure 1: Performance comparison across approaches: Our two-stage pipeline substantially outperforms direct methods on heterogeneous clinical data.

with CAD-RADS scores. The dataset spans a four-year period and reflects substantial variation in clinical language, terminology, and reporting style across multiple cardiologists. Characteristics of the proposed dataset are summarized in Table 1.

Furthermore, the direct LLM approaches (Wang et al., 2024; Arnold et al., 2025; Min et al., 2025) underperform when evaluated on clinical data exhibiting substantial variation in language and report structure. A possible reason is that LLMs struggle to jointly handle linguistic heterogeneity and clinical reasoning. We propose a two-stage approach that decouples structuring from classification: an LLM-based parser normalizes heterogeneous reports into structured JSON format, followed by fine-tuned BERT classification. As shown in Figure 1, the proposed two-stage pipeline improves the F1 score by 6%-13% compared with direct BERT models or direct LLM methods. Our key contributions are:

- We introduce **CCTA-RADS**, the largest publicly available, real-world dataset for CAD-RADS scoring from clinical text, enabling reproducible research in this domain.
- We demonstrate that **direct state-of-the-art approaches fail** on heterogeneous, real-world reports. We propose and validate a **two-stage pipeline** that handles report heterogeneity and achieves substantial performance improvements across all evaluated models.
- We deploy our system as an interactive web interface that allows clinicians to upload CCTA reports for automated CAD-RADS assessment with explainability visualizations.
- We provide comprehensive analysis and release the dataset, structured representations, code, and trained models to ensure **reproducibility** at https://github.com/bsudharshan2001/cctarads.

Study	Size	Type	Availability	Approach
(Wang et al., 2024)	590	Real	D×C×W×	Direct LLM
(Arnold et al., 2025)	200	Synth.	D×C×W×	Direct LLM
(Min et al., 2025)	319	Real	D✓C×W×	Direct LLM
Ours	**940**	**Real**	D✓ C✓ W✓	**Two-stage**

Table 1: Comparison of CAD-RADS classification approaches. D, C, and W stand for dataset, code, and model weights.

2 Related Work

2.1 Clinical Text Datasets

Several clinical text datasets have been developed for NLP research. General radiology datasets include MIMIC-CXR (Johnson et al., 2019) with over 220,000 chest X-ray reports, RadGraph (Jain et al., 2021) with chest X-ray annotations, and TECRR (Hussain et al., 2024) with breast imaging reports using BI-RADS categories. While valuable for their respective domains, these datasets focus on different imaging modalities and scoring systems.

For CCTA analysis, existing image datasets like ImageCAS (Zeng et al., 2023), CorArtTS 2020 (Wang et al., 2023), and CCA-200 (Yang et al., 2025) focus on coronary artery segmentation from imaging data rather than textual analysis. These resources facilitate image processing tasks but do not provide textual reports for automated CAD-RADS scoring.

2.2 CAD-RADS Classification from Text

Recent pioneering work has explored automated CAD-RADS scoring directly from clinical text. Wang et al. (Wang et al., 2024) evaluated GPT-3.5, GPT-4, and Llama3 on 590 cardiac CT reports, achieving good agreement with radiologists (AC1 = 0.861-0.941). Arnold et al. (Arnold et al., 2025) tested multiple LLMs on 200 cardiac CT reports, with GPT-4o and Llama3 70b achieving 93% and 92.5% accuracy respectively. Similarly, Min et al. (Min et al., 2025) evaluated GPT-4 and other LLMs on 319 cardiac CT reports, achieving comparable performance. These studies demonstrate the feasibility of LLM-based CAD-RADS classification and establish important baselines for the field. Table 1 summarizes the key differences between these

approaches and our work.

However, these approaches face critical limitations. **First, dataset constraints** limit their impact: studies use small datasets (200-590 reports) and none release their data publicly, preventing reproducibility. Arnold et al. (Arnold et al., 2025) acknowledge using synthetically augmented data, while Min et al. (Min et al., 2025) use only 319 reports, raising questions about whether reported performance generalizes to authentic clinical documentation at scale.

Second, methodological assumptions remain unvalidated. All three approaches rely on direct-to-LLM classification, assuming models can handle clinical report heterogeneity without structural preprocessing. This assumption has not been validated against large-scale, heterogeneous data, and none provide failure analysis addressing inconsistent reporting styles, varied terminology, and structural variations in real clinical practice.

Our work addresses these gaps by providing a large-scale, public, real-world dataset enabling systematic evaluation. We demonstrate that structured preprocessing significantly improves performance over direct methods on heterogeneous clinical data.

3 CCTA-RADS Dataset

3.1 Data Collection, Annotation, and De-identification

The CCTA-RADS dataset comprises 940 real-world CCTA reports collected from Tampa General Hospital (TGH) imaging department (2020-2024) [1] under IRB approval. Reports were authored by board-certified radiologists and cardiologists during routine care, with CAD-RADS scores assigned by expert clinicians following standard guidelines. The dataset includes reports annotated using both CAD-RADS 1.0 and 2.0; the core scoring categories (0-5, N) remain identical between versions (see Appendix A for versioning details). Most report fields are pre-populated through automated systems, with final CAD-RADS scores validated by interpreting radiologists. All reports underwent HIPAA-compliant de-identification by certified specialists.

3.2 Dataset Statistics and Format

Table 2 summarizes the dataset characteristics and label distribution. The dataset exhibits natural class imbalance reflecting real clinical practice, with CAD-RADS 0 (no stenosis) being most common and CAD-RADS 5 (total occlusion) being rarest.

Dataset Statistics		Label Distribution	
Reports	940	0	338 (35.96%)
Avg length	529	1	140 (14.89%)
Median length	518	2	134 (14.26%)
Min length	309	3	74 (7.87%)
Max length	1020	4	72 (7.66%)
Vocabulary	6798	5	7 (0.74%)
Avg sentences	42	N	175 (18.62%)

Table 2: Dataset statistics and CAD-RADS label distribution.

The reports in our dataset exhibit considerable variation in length and detail, reflecting the diversity of clinical documentation practices and the varying complexity of cases. The vocabulary includes specialized cardiac and radiological terminology, with a significant number of domain-specific terms related to coronary anatomy, pathology, and imaging findings.

For instance, stenosis information might be presented in dedicated 'Findings' subsections, embedded within overall impressions, or use varied phrasing for similar severities. The CCTA-RADS dataset encapsulates this real-world linguistic diversity, making it a robust benchmark for NLP models designed to handle such variations. Table 3 provides illustrative examples of common linguistic and structural variations encountered in the dataset, drawn from actual reports. Our preprocessing pipeline (Section 4.2) is designed to normalize many of these structural and terminological differences into a structured format.

4 Experimental Evaluation

The heterogeneous nature of real-world CCTA reports presents significant challenges for automated CAD-RADS classification. To systematically evaluate different approaches, we first established baselines using direct methods similar to recent work (Wang et al., 2024; Arnold et al., 2025; Min et al., 2025), then developed our two-stage pipeline to address their limitations. This section presents our comprehensive evaluation revealing why direct approaches fail on realistic clinical data and how structured preprocessing enables robust classification.

[1] https://www.tgh.org/institutes-and-services/imaging

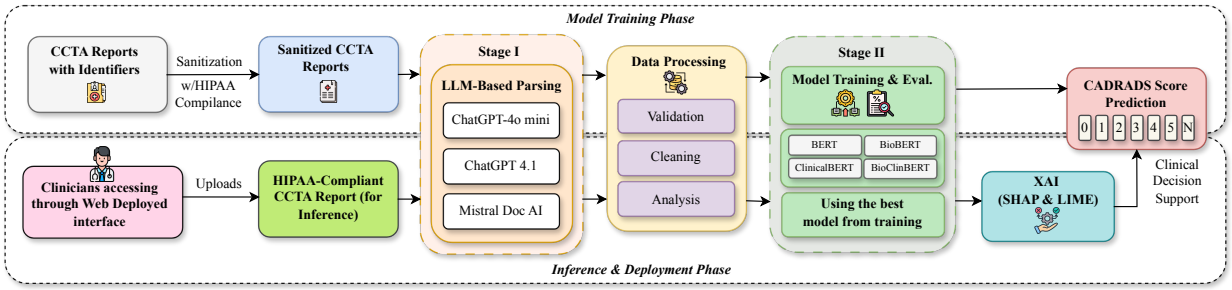

Figure 2: System pipeline for automated CAD-RADS score prediction, showing the flow from sanitized CCTA reports through LLM-based parsing, data processing, and model training to final prediction.

Aspect of Variation	Excerpt from Report 5 (CAD-RADS 1 P1)	Excerpt from Report 9 (CAD-RADS 1 P1)
Stenosis Description (LCx - Findings)	"The LCx was poorly opacified, but appears grossly patent with calcified plaque in the proximal segment resulting in less than 25% luminal stenosis."	"The left circumflex artery is patent with <24% calcified plaque proximally."
Stenosis Description (LAD - Findings)	"The left anterior descending artery was poorly opacified, but appears grossly patent without evidence of plaque or stenosis."	"Less than 24% low attenuation plaque with flecks of calcification in proximal LAD."
Impression - Stenosis Summary	"* There is calcified plaque in the proximal left circumflex coronary artery resulting in less than 25% luminal stenosis."	"* Less than 24% low attenuation plaque with flecks of calcification in proximal LAD. * Less than 24% low attenuation plaques in the proximal RCA."

Table 3: Direct comparison of variations in excerpts from two reports with the same CAD-RADS score (both CAD-RADS 1 P1).

4.1 The Challenge: Failure of Direct Classification Approaches

We first evaluated direct classification approaches to establish baseline performance and understand the challenges posed by heterogeneous clinical reports. Our task is multi-class classification: given a de-identified CCTA report, predict its CAD-RADS score (0-5 or N). Evaluation metrics are defined in Section 5.

4.1.1 Direct LLM Classification

Following recent work (Wang et al., 2024; Arnold et al., 2025; Min et al., 2025), we evaluated state-of-the-art LLMs using various prompting strategies. We tested GPT-4o and GPT-o3 (both non-reasoning and reasoning variants) with zero-shot and one-shot (Brown et al., 2020) and Chain-of-Thought (CoT) (Wei et al., 2023) approaches on our dataset (Table 4).

While LLMs achieved reasonable overall performance (F1 82-84%), the class-wise results reveal significant limitations. Most critically, LLMs struggled with CAD-RADS N (non-diagnostic), achieving only 42-54% accuracy. LLMs failed to reliably identify subtle indicators of poor image quality or technical limitations that characterize non-diagnostic studies.

4.1.2 Fine-Tuning on Raw Text

We also evaluated BERT-family models (Devlin et al., 2019; Lee et al., 2020; Huang et al., 2020; Alsentzer et al., 2019) fine-tuned directly on raw, unprocessed report text to assess whether domain-adapted encoders could handle clinical heterogeneity without explicit structuring (Table 5).

Raw text fine-tuning yielded substantially worse performance (F1 70-77%) compared to our final approach (90.12%). The class-wise results reveal severe limitations: models showed highly inconsistent performance across clinically critical classes, with particularly poor reliability for intermediate CAD-RADS categories. BioBERT paradoxically performed worst overall despite being domain-adapted, highlighting how raw clinical text overwhelms even specialized models without proper structuring.

The failure of both direct LLM classification and raw text fine-tuning can be attributed to three fundamental challenges: **(1) Simultaneous cognitive load:** These approaches force models to simultaneously handle linguistic heterogeneity (parsing diverse formats, terminology, and structures) and clinical reasoning (understanding stenosis severity and diagnostic implications), creating competing optimization objectives. **(2) Format noise over-**

Approach	Model	Overall Performance (%)				Class-wise Accuracy (%)						
		Acc	Prec	Rec	F1	0	1	2	3	4	5	N
Zero Shot	GPT-4o	84.50	84.15	84.82	82.76	97.59	92.72	95.14	94.19	92.22	76.47	43.40
	o3	85.71	83.78	85.70	83.49	97.05	95.36	95.14	93.02	92.22	75.47	50.65
One Shot	GPT-4o	83.78	82.28	84.00	81.32	97.85	92.00	95.80	95.29	89.89	75.00	42.17
	o3	85.35	83.54	86.92	83.90	96.51	94.67	95.10	92.94	92.13	87.50	49.57
CoT	GPT-4o	84.62	82.14	84.64	81.85	97.59	92.72	95.83	95.35	87.78	76.47	46.75
	o3	86.07	83.85	86.51	84.08	96.78	94.70	93.75	91.86	92.22	82.35	53.91

Table 4: Direct LLM classification: overall performance and class-wise accuracy.

Model	Overall Performance (%)				Class-wise Accuracy (%)						
	Acc	Prec	Rec	F1	0	1	2	3	4	5	N
BERT	75.00	74.21	75.00	73.00	95.92	83.87	70.09	93.75	0.00	0.00	58.13
BioBERT	71.95	70.80	71.95	70.77	83.67	83.87	45.00	31.25	41.14	0.00	91.67
ClinicalBERT	76.83	77.39	76.83	76.74	87.76	77.42	65.00	62.50	66.67	0.00	83.33
Bio+ClinicalBERT	77.44	81.09	77.44	77.31	97.96	74.19	85.00	56.25	88.89	0.00	61.11

Table 5: BERT models on raw text: overall performance and class-wise accuracy.

whelming signal: The substantial variations in section headers, terminology, and report structures create noise that obscures the underlying clinical patterns, making it difficult for models to learn consistent decision boundaries. **(3) Inconsistent feature extraction:** Without standardized input representation, models must learn to extract clinically relevant features from inconsistently formatted text, leading to unreliable feature representations that vary based on reporting style rather than clinical content. These challenges suggest that report structuring and clinical classification should be treated as separate tasks.

4.2 A Two-Stage Pipeline for Robust Classification

Given the failure of direct methods on heterogeneous clinical data, we developed a two-stage pipeline that decouples report structuring from classification (Figure 2). Our key insight is that rather than forcing a single model to handle both linguistic variability and clinical reasoning, we can leverage different model strengths: LLMs excel at understanding and parsing diverse text structures, while fine-tuned BERT models excel at classification on clean, structured data.

4.2.1 Stage 1: LLM-Powered Report Structuring

We employed ChatGPT-4o-mini to parse raw DOCX reports into standardized JSON format. This stage handles the three main sources of heterogeneity:

1. **Section Normalization:** Identifying semantically equivalent sections ("FINDINGS" vs "RESULTS") regardless of formatting
2. **Clinical Entity Extraction:** Extracting vessel assessments, stenosis percentages, and technical quality indicators from varied locations and phrasings
3. **Terminology Standardization:** Normalizing diverse expressions ("25% stenosis" vs "mild narrowing") into consistent representations

The LLM parsing stage focuses exclusively on text normalization without making classification decisions, separating heterogeneous text handling from clinical reasoning. To validate parser robustness, we compared outputs on reports with and without existing CAD-RADS scores; BERTScore F1 (97.47%) demonstrated consistent structuring regardless of pre-existing scores. Cross-model validation showed high semantic consistency across LLM architectures (BERTScore F1 > 0.93; Table 6). The pipeline underwent validation through multiple iterations with practicing cardiologists and cardiovascular imaging specialists, ensuring clinical accuracy across diverse reporting styles.

	GPT-4o	Mistral	GPT-4.1
Original	0.836	0.835	0.846
GPT-4o	**1.000**	0.931	0.934
Mistral	0.931	**1.000**	0.944
GPT-4.1	0.934	0.944	**1.000**

Table 6: BERTScore F1 comparison matrix between LLM parsers and original reports.

4.2.2 Stage 2: Fine-Tuning on Structured Data

The structured JSON output feeds into fine-tuned BERT models for classification. By operating on normalized, structured text, these models can focus on clinical reasoning rather than parsing heterogeneous formatting. The two-stage separation allows each component to leverage its strengths while avoiding their respective weaknesses. To address class imbalance, particularly for rare CAD-RADS 5 cases, we employed inverse frequency class weighting during training (details in Appendix B.2).

5 Results and Analysis

5.1 Experimental Setup

We used consistent experimental settings across all approaches for fair comparison. The dataset was split into train (70%), validation (15%), and test (15%) sets. We used Optuna for hyperparameter optimization (20 trials per model), tuning learning rate, epochs, batch size, and sequence length. We report accuracy, precision, recall, and F1 score.

5.2 Classification Performance

Our two-stage pipeline consistently improves performance over direct approaches. Table 7 shows the performance of BERT-family models on structured data, compared to the baseline results in Tables 4 and 5.

The two-stage pipeline substantially improves performance across all models compared to direct approaches. All models achieve better results when operating on structured data, with BioBERT reaching the highest F1 score of 90.12%, compared to 84.08% for direct LLM prompting and 77.31% for raw-text fine-tuning. This consistent improvement across all models validates our core hypothesis that structured preprocessing is critical for handling heterogeneous clinical text.

The success of our two-stage approach stems from three key design principles: **(1) Separation of concerns:** By decoupling text normalization from clinical classification, each model can focus on its core strength: LLMs excel at understanding and parsing diverse text structures, while fine-tuned BERT models excel at pattern recognition on clean, structured data. **(2) Noise elimination:** The LLM parser removes formatting inconsistencies, terminology variations, and structural heterogeneity, allowing the classifier to focus purely on clinical reasoning without being distracted by presentation differences. **(3) Consistent feature representation:** Structured JSON output ensures that all reports are represented in a standardized format, enabling the classifier to learn robust clinical patterns based on content rather than formatting style.

The class-wise results demonstrate clinical value: all models achieve perfect accuracy (100%) on CAD-RADS 3 and 4 (moderate to severe stenosis), ensuring reliable high-risk patient identification. The structured approach improves performance on challenging classes, with CAD-RADS N accuracy reaching 68-73% compared to 42-54% for direct LLM approaches.

5.3 Error Analysis and Explainability

The confusion matrix (Table 8) reveals errors concentrated in CAD-RADS N (often misclassified as 0 or 1) and the rare CAD-RADS 5 class, with misclassifications occurring between adjacent categories consistent with ordinal CAD-RADS structure. In medical scoring, misclassifying a level 1 as a level 0 is far better than misclassifying a level 4 as a level 0 (a catastrophic error). We further utilize SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016) to highlight the importance of report components in making a decision. Figure 3 shows aggregate SHAP-based section importance scores for correctly classified CAD-RADS N reports: 'Technical Quality' dominates predictions (0.516), followed by 'Impression' (0.181). Instance-level and LIME visualizations are provided in Appendix B.5.

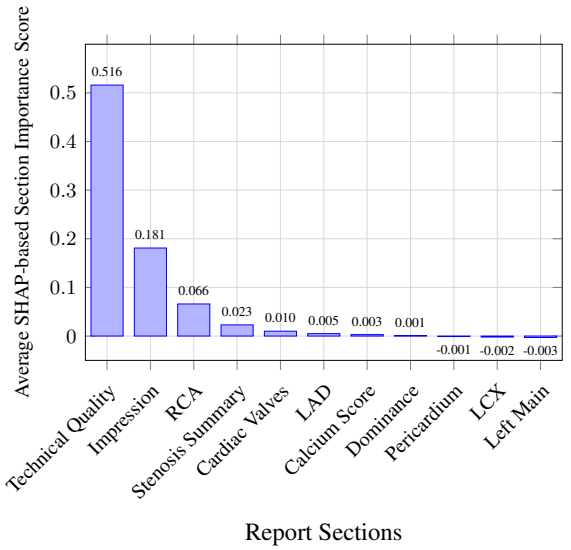

Figure 3: Aggregate SHAP-based section importance scores for correctly classified CAD-RADS N reports.

Model	Overall Performance (%)				Class-wise Accuracy (%)						
	Acc	Prec	Rec	F1	0	1	2	3	4	5	N
Direct LLM (o3)	*86.07*	*83.85*	*86.51*	*84.08*	*96.78*	*94.70*	*93.75*	*91.86*	*92.22*	*82.35*	*53.91*
Raw Text (Bio+Clin.)	*77.44*	*81.09*	*77.44*	*77.31*	*97.96*	*74.19*	*85.00*	*56.25*	*88.89*	*0.00*	*61.11*
BERT	86.52	84.37	83.91	84.08	96.15	91.30	87.50	100.00	100.00	0.00	72.73
ClinicalBERT	89.36	87.24	86.53	86.82	96.15	91.30	83.33	100.00	100.00	0.00	72.73
BioBERT	**92.25**	**90.42**	**89.87**	**90.12**	**100.00**	**95.65**	**91.67**	**100.00**	**100.00**	**0.00**	**68.18**
Bio+ClinicalBERT	91.48	89.75	88.92	89.31	100.00	95.65	79.17	100.00	100.00	0.00	72.73

Table 7: Two-stage pipeline: overall performance and class-wise accuracy.

	0	1	2	3	4	5	N
0	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
1	4.3%	95.7%	0.0%	0.0%	0.0%	0.0%	0.0%
2	0.0%	4.2%	91.7%	4.2%	0.0%	0.0%	0.0%
3	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
4	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
5	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
N	9.1%	13.6%	4.5%	4.5%	0.0%	0.0%	68.2%

Table 8: Confusion matrix (%) of BioBERT predictions on the test set. Rows: true labels, Columns: prediction.

5.4 Web-based Deployment

Our trained model is deployed as an interactive web interface (Figure 4) that allows clinicians to upload CCTA reports for automated CAD-RADS assessment with SHAP and LIME explainability visualizations. Details are in Section B.6. A demo video of the system is available in our code repository.

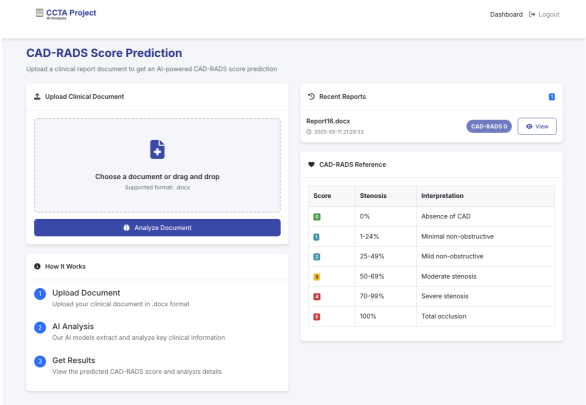

Figure 4: Screenshot of the web interface showing the report upload page with instructions for clinicians.

6 Discussion and Conclusion

6.1 Discussion

Our two-stage pipeline achieved 90.12% F1 with BioBERT, improving over direct LLM (84.08%) and raw-text fine-tuning (77.31%), demonstrating that **structuring clinical text is critical for high-performance classification** on heterogeneous documentation. The clinical implications are significant: perfect accuracy on CAD-RADS 3-4 ensures reliable high-risk patient identification, while improved CAD-RADS N classification (68-73% vs 42-54%) demonstrates effective leveraging of technical quality indicators. Such systems could standardize CAD-RADS assignment and reduce inter-reader variability.

6.2 Limitations

Key limitations include the dataset's single-institution origin, which may limit generalizability to inter-institutional heterogeneity and requires external validation. Evaluating model robustness under such cross-institutional variation will be an important direction for future work. Class imbalance remains challenging, particularly for CAD-RADS 5 (only 7 examples). While class weighting helps, it has inherent limitations for extremely rare classes. Exploring a few-shot learning or data synthesis strategy to better address rare categories is left for future work.

6.3 Conclusion

We introduced CCTA-RADS, the largest public dataset of real-world CCTA reports (940 reports), capturing heterogeneity in clinical terminology and reporting style. We have demonstrated that standard approaches struggle with heterogeneous clinical data and proposed a two-stage pipeline. By leveraging LLM parsing before BERT classification, the proposed approach provides substantial performance improvements and establishes a reproducible benchmark for clinical text analysis.

Acknowledgments

This paper is supported by the Office of Naval Research subawarded through Virginia Tech under Grant N00014-24-1-2730.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Philipp Georg Arnold, Maximilian Frederik Russe, Fabian Bamberg, Tilman Emrich, Milán Vecsey-Nagy, Ayaat Ashi, Dmitrij Kravchenko, Ákos Varga-Szemes, Martin Soschynski, Alexander Rau, Elmar Kotter, and Muhammad Taha Hagar. 2025. Performance of large language models for cad-rads 2.0 classification derived from cardiac ct reports. *Journal of Cardiovascular Computed Tomography*, 19(3):322–330.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners.

Tianrun Cai, Andreas A. Giannopoulos, Sheng Yu, Tatiana Kelil, Beth Ripley, Kanako K. Kumamaru, Frank J. Rybicki, and Dimitrios Mitsouras. 2016. Natural language processing technologies in radiology research and clinical applications. *RadioGraphics*, 36(1):176–191.

Ricardo C. Cury, Suhny Abbara, Stephan Achenbach, Arthur Agatston, Daniel S. Berman, Matthew J. Budoff, Karin E. Dill, Jill E. Jacobs, Christopher D. Maroules, Geoffrey D. Rubin, Frank J. Rybicki, U. Joseph Schoepf, Leslee J. Shaw, Arthur E. Stillman, Charles S. White, Pamela K. Woodard, and Jonathon A. Leipsic. 2016. CAD-RADS™: Coronary Artery Disease - Reporting and Data System: An Expert Consensus Document of the Society of Cardiovascular Computed Tomography (SCCT), the American College of Cardiology (ACC), and the American College of Radiology (ACR). *JACC: Cardiovascular Imaging*, 9(9):1099–1113.

Ricardo C. Cury, Jonathon Leipsic, Suhny Abbara, Stephan Achenbach, Daniel Berman, Marcio Bittencourt, Matthew Budoff, Kavitha Chinnaiyan, Andrew D. Choi, Brian Ghoshhajra, Jill Jacobs, Lynne Koweek, John Lesser, Christopher Maroules, Geoffrey D. Rubin, Frank J. Rybicki, Leslee J. Shaw, Michelle C. Williams, Eric Williamson, and 3 others. 2022. CAD-RADS™ 2.0 - 2022 Coronary Artery Disease - Reporting and Data System: An Expert Consensus Document of the Society of Cardiovascular Computed Tomography (SCCT), the American College of Cardiology (ACC), the American College of Radiology (ACR) and the North America Society of Cardiovascular Imaging (NASCI). *JACC: Cardiovascular Imaging*, 15(11):1974–2001.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Elizabeth Ford, John A Carroll, Helen E Smith, Donia Scott, and Jackie A Cassell. 2016. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 23(5):1007–1015.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission.

Sadam Hussain, Usman Naseem, Mansoor Ali, Daly Betzabeth Avendaño Avalos, Servando Cardona-Huerta, Beatriz Alejandra Bosques Palomo, and Jose Gerardo Tamez-Peña. 2024. TECRR: a benchmark dataset of radiological reports for BI-RADS classification with machine learning, deep learning, and large language model baselines. *BMC Medical Informatics and Decision Making*, 24(1):310.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. volume 36, pages 1234–1240.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions.

S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. 2008. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of Medical Informatics*, 17(1):128–144.

Dabin Min, Kwang Nam Jin, SangHeum Bang, Moon Young Kim, Hack-Lyoung Kim, Won Gi Jeong, Hye-Jeong Lee, Kyongmin Sarah Beck, Sung Ho Hwang, Eun Young Kim, and Chang Min Park. 2025. Large language models for CAD-RADS

2.0 extraction from semi-structured coronary CT angiography reports: A multi-institutional study. *Korean Journal of Radiology*, 26(9):817–831. PMID: 40873373; PMCID: PMC12394816.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier.

Qianjin Wang, Lisheng Xu, Lu Wang, Xiaofan Yang, Yu Sun, Benqiang Yang, and Stephen E. Greenwald. 2023. Automatic coronary artery segmentation of ccta images using unet with a local contextual transformer. *Frontiers in Physiology*, Volume 14 - 2023.

Yuli Wang, Wen-Chi Hsu, Yuwei Dai, Victoria Shi, Gigin Lin, Harrison Bai, and Cheng Ting Lin. 2024. Automatic assignment of cad-rads categories in coronary cta reports using large language model. *Circulation*, 150(Suppl_1):A4119869–A4119869.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Xiaoyu Yang, Lijian Xu, Simon Yu, Qing Xia, Hongsheng Li, and Shaoting Zhang. 2025. Segmentation and vascular vectorization for coronary artery by geometry-based cascaded neural network. *IEEE Transactions on Medical Imaging*, 44(1):259–269.

An Zeng, Chunbiao Wu, Guisen Lin, Wen Xie, Jin Hong, Meiping Huang, Jian Zhuang, Shanshan Bi, Dan Pan, Najeeb Ullah, Kaleem Nawaz Khan, Tianchen Wang, Yiyu Shi, Xiaomeng Li, and Xiaowei Xu. 2023. Imagecas: A large-scale dataset and benchmark for coronary artery segmentation based on computed tomography angiography images. *Computerized Medical Imaging and Graphics*, 109:102287.

Supplementary Materials

A Dataset Details

A.1 CAD-RADS Classification System

Table 9 provides the complete CAD-RADS classification system referenced in the main paper.

A.2 Dataset Schema

Table 10 provides the complete schema for the structured JSON format of the CCTA-RADS dataset.

A.3 Detailed Annotation Methodology

The CAD-RADS scores were assigned through a multi-step process ensuring clinical accuracy:

Scores were assigned by certified radiologists, cardiologists, and clinicians as part of routine clinical care. Most report fields are pre-populated through automated systems (calcium scores, artery names, automated stenosis labels), with the final CAD-RADS score manually validated and documented by the interpreting radiologist.

For reports generated before the widespread adoption of CAD-RADS 2.0 or lacking explicit scores, retrospective annotation was performed by cardiovascular imaging specialists following standardized guidelines. As detailed in the main paper, our dataset contains reports annotated using both CAD-RADS 1.0 (2020-2022 reports) and CAD-RADS 2.0 (2022-2024 reports) guidelines. The core scoring categories (0-5, N) for maximal stenosis degree classification remain identical between versions, which is the primary factor our models classify.

B Dataset Analysis

B.1 Label Distribution

This class imbalance presents a significant challenge for machine learning models, particularly for the rare but clinically critical CAD-RADS 5 category. The CAD-RADS N category (non-diagnostic studies) represents 18.7% of the dataset, highlighting another important aspect of real-world clinical practice where image quality or other factors may limit diagnostic certainty.

B.2 Class Imbalance Mitigation

To address class imbalance in the dataset, particularly for underrepresented categories like CAD-RADS 5, we employed class weighting during training to penalize misclassifications of minority classes more heavily. Class weights were calculated using inverse frequency weighting to ensure balanced learning across all CAD-RADS categories.

Specifically, for each class i, the weight w_i was computed as:

$$w_i = \frac{N}{n_i \cdot C} \tag{1}$$

where N is the total number of training samples, n_i is the number of samples in class i, and C is the total number of classes. These weights were then applied to the cross-entropy loss function during training:

$$\mathcal{L} = -\sum_{i=1}^{C} w_i \cdot y_i \log(\hat{y}_i) \tag{2}$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability for class i. This approach helps im-

CAD-RADS Categories			CAD-RADS Modifiers	
Score	**Stenosis Severity**	**Clinical Implication**	**Modifier**	**Description**
0	0% (Absent)	Very low likelihood of CAD	V	Presence of vulnerable plaque
1	1-24% (Minimal)	Low likelihood of CAD	S	Presence of stent
2	25-49% (Mild)	Low-to-intermediate likelihood of CAD	G	Presence of bypass graft
3	50-69% (Moderate)	Intermediate likelihood of CAD	P1	Plaque burden: minimal
4	70-99% (Severe)	High likelihood of CAD	P2	Plaque burden: mild
5	100% (Total occlusion)	Very high likelihood of CAD	P3	Plaque burden: moderate
N	Non-diagnostic	Additional evaluation needed	P4	Plaque burden: extensive

Table 9: Complete CAD-RADS classification system showing stenosis categories and modifiers.

Field	Description
report_id	Anonymized unique identifier for each report
history	Clinical indication and patient history
technique	Technical parameters of imaging procedure
contrast	Details of contrast agent used
medications	Medications administered during procedure
dose	Radiation dose information
technical_quality	Assessment of image quality
calcium_score	Calcium scores for coronary arteries (LM, LAD, LCX, RCA, PDA)
findings	Vessel-by-vessel assessment of stenosis and plaque
impression	Summary of key findings and interpretation
cadrads_score	Primary CAD-RADS category (0-5, N)
cadrads_modifiers	Additional modifiers (V, S, G, P1-P4)

Table 10: Complete dataset schema for the structured JSON format.

prove model learning capacity for critical high-risk cases despite limited training examples.

B.3 HIPAA-defined identifiers

All reports underwent rigorous de-identification following HIPAA Safe Harbor guidelines. The de-identification process involved manual removal of all 18 HIPAA-defined identifiers by certified specialists:

Some of the HIPAA-defined identifiers include:

- Names of patients and relatives
- Dates (except year) related to the patient
- Medical record numbers and other identifying numbers
- Geographic information more specific than state
- Any other unique identifying characteristics

B.4 LIME Explainability Visualization

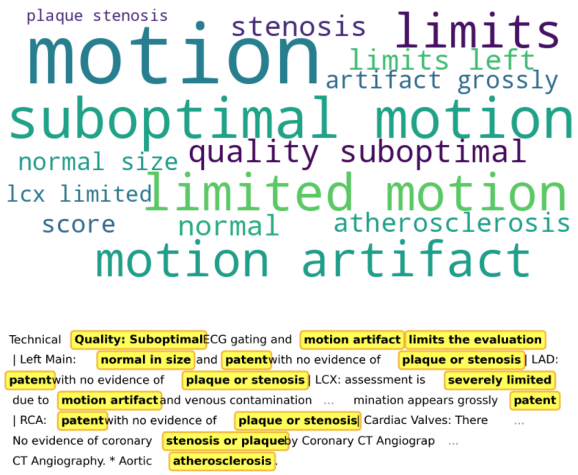

Figure 5: LIME explainability visualization for a representative, correctly classified CAD-RADS N report. **Top:** Word cloud showing the most influential words for the model's prediction. **Bottom:** The actual report text with important words highlighted.

B.5 Instance-Level SHAP-based Section Analysis

To provide a more granular understanding of how different report sections influence the BioBERT model's predictions for individual cases, we present SHAP-based section importance analysis results for specific instances. SHAP-based section importance analysis was performed by masking individual report sections and observing the change in prediction probability for the target class. Figures 6 and 7 illustrate this for two CAD-RADS N reports: one correctly classified by our best model, BioBERT, and one misclassified.

Instance-level SHAP-based section importance analysis showing section-wise contributions to model predictions is shown for two example cases.

Figure 6 displays the instance-level SHAP-based section importance scores for a CCTA report that was correctly classified as CAD-RADS N. The 'Technical Quality' section demonstrates a dom-

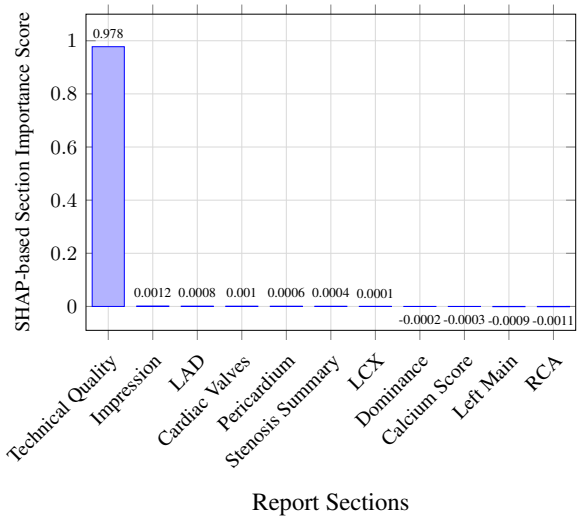

Figure 6: Instance-level SHAP-based section importance scores for a correctly classified CAD-RADS N report

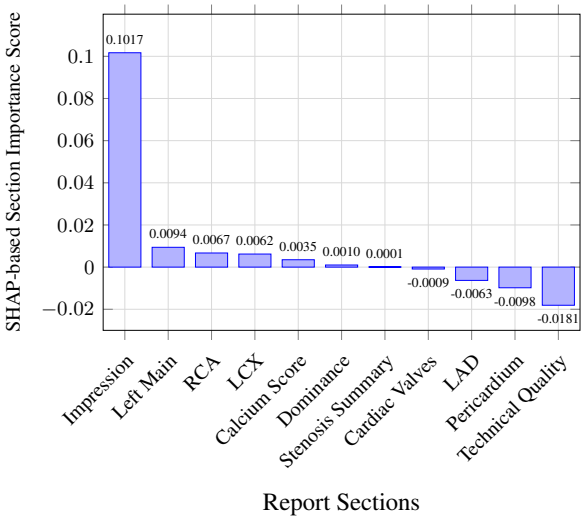

Figure 7: Instance-level SHAP-based section importance scores for a misclassified CAD-RADS N report (predicted as CAD-RADS 4)

inant positive importance score, signifying its crucial role in the model's correct identification of this non-diagnostic case. This aligns with our aggregate SHAP-based section importance findings (see Figure 3 in the main paper), where 'Technical Quality' was also found to be the most influential section on average for CAD-RADS N predictions. Other sections, such as 'Impression' and 'LAD', show minimal positive contributions for this specific correct prediction, further emphasizing the model's reliance on direct cues from the 'Technical Quality' section.

In contrast, Figure 7 illustrates the instance-level SHAP-based section importance scores for a CAD-RADS N report that was misclassified by BioBERT as CAD-RADS 4. For this incorrect prediction towards CAD-RADS 4, the 'Impression' section has the highest positive SHAP-based section importance score, suggesting that content within the impression (likely describing findings that could be interpreted as indicative of significant disease) strongly pushed the model towards the erroneous CAD-RADS 4 classification. Other sections detailing specific coronary arteries like 'Left Main', 'RCA', and 'LCX' also contributed positively, albeit to a lesser extent, to this misprediction. Intriguingly, the 'Technical Quality' section shows a negative SHAP-based section importance score for the CAD-RADS 4 prediction. This indicates that the content of the 'Technical Quality' section in this particular report was actually arguing against a CAD-RADS 4 classification (and potentially contained signals indicative of CAD-RADS N), but these signals were ultimately overridden by the stronger perceived pathological indicators in other sections, notably the 'Impression'.

These instance-level SHAP-based section analyses complement the aggregate SHAP-based section importance scores by demonstrating the variability in section contributions from report to report. They offer valuable insights into specific model decision pathways and potential reasons for misclassification, such as when strong but misleading signals in one section (e.g., 'Impression') overpower weaker, albeit correct, signals in another section (e.g., 'Technical Quality' for an N class).

B.6 Web-based Deployment

The web application allows clinicians to upload CCTA reports directly through a user-friendly interface. Upon submission, the system processes the report using our LLM-based parsing pipeline to extract structured information from the free-text report. This extracted data is then fed into our fine-tuned BioBERT model to generate an automated CAD-RADS prediction, which is then presented to the clinician along with an interpretation guide for the score (Figure 8).

To enhance transparency and clinical trust in the automated predictions, the web interface provides explainable AI visualizations for each prediction. Figure 9 displays a LIME-based explanation showing which specific words and phrases in the uploaded report most strongly influenced the model's CAD-RADS classification decision. This word-level attribution helps clinicians understand

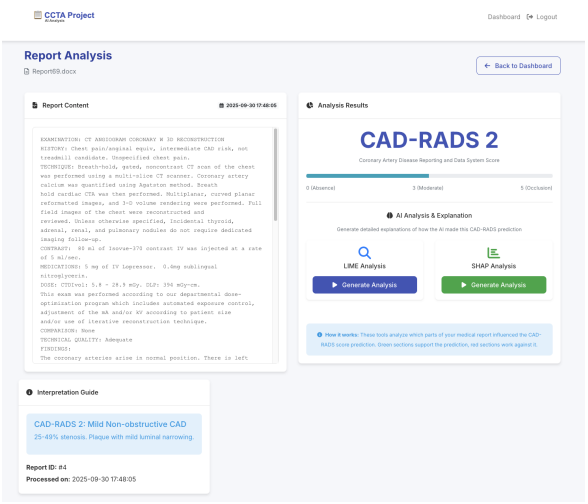

Figure 8: Screenshot of the results page showing the predicted CAD-RADS score, extracted findings, and confidence metrics for clinical review.

the reasoning behind the automated assessment and identify key clinical indicators that drove the prediction.

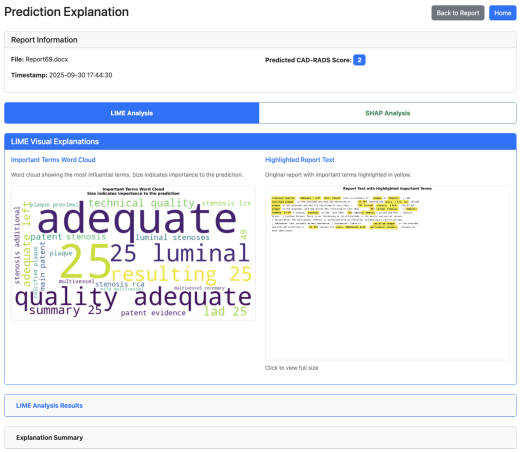

Figure 9: LIME-based prediction explanation in the web interface, highlighting influential words and phrases that contributed to the CAD-RADS classification.

Additionally, the interface presents SHAP-based section importance analysis (Figure 10), which quantifies the relative contribution of each standardized report section (e.g., Technical Quality, Impression, vessel-specific findings) to the final prediction. This section-level explanation complements the word-level LIME analysis by providing a higher-level view of which parts of the clinical report were most diagnostically relevant for the automated CAD-RADS assessment.

This deployment represents a practical implementation of our research, demonstrating how NLP

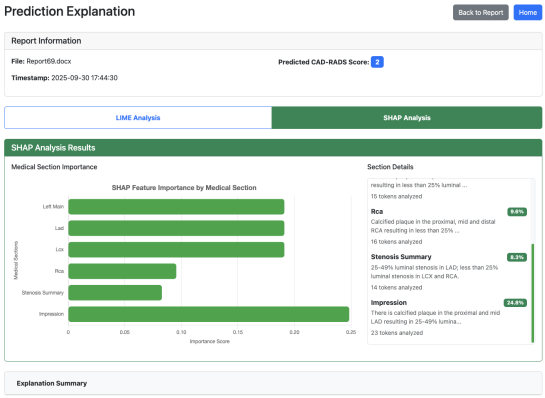

Figure 10: SHAP-based section importance analysis in the web interface, showing the relative contribution of report sections to the predicted CAD-RADS score.

models trained on the CCTA-RADS dataset can be integrated into clinical workflows to provide decision support. The system is designed to assist, not replace, expert radiologist assessment, offering a preliminary CAD-RADS classification that can help standardize reporting and potentially reduce inter-reader variability. The integration of explainable AI features ensures that clinicians can validate and understand the automated predictions, promoting appropriate trust and facilitating clinical adoption.

B.7 LLM-based Parsing Prompt

To address the challenge of structural and terminological heterogeneity in CCTA reports, we developed a specialized prompt for our LLM-based parsing pipeline. The prompt (see Table 11) instructs the model to extract standardized information from reports with varying formats. The extracted scores and modifiers from training data undergo cross-validation to ensure accuracy before being used for model training.

This prompt enables consistent extraction of CAD-RADS-relevant information despite variations in report structure, terminology, and formatting across different clinicians and institutions. The structured JSON output provides a normalized representation for downstream classification tasks.

LLM Parsing Instructions

Given the following medical report, extract the relevant features for predicting the CADRADS score and convert them into JSON format with the following structure. Ignore any subheadings or sections not specified below. If any of the required fields are missing in the report, set their values to an empty string or null. Ensure that the output JSON strictly follows the structure provided, without including any additional or random information from the report.

```
{
    "left_main": "...",
    "lad": "...",
    "lcx": "...",
    "rca": "...",
    "stenosis_summary": "...",
    "calcium_score": "...",
    "cardiac_valves": "...",
    "pericardium": "...",
    "dominance": "...",
    "technical_quality": "...",
    "impression": "...",
    "cadrads_score": "...",
    "modifiers": ["..."]
}
```

1. **Extract only the information corresponding to the fields specified in the JSON structure.**
2. **Ignore any random or additional subheadings and their content not defined in the JSON structure.**
3. **For missing sections in the report, use an empty string `""` or `null` in the JSON output.**
4. **Do not include any extra fields or data not specified in the JSON structure.**
5. **IMPORTANT: DO NOT HALLUCINATE OR INFER VALUES. Only extract information explicitly stated in the report.**
   - If the CADRADS score is not explicitly mentioned in the report, use `"N"` (not present).
   - If features like vessel conditions or calcium scores are not mentioned, use empty strings rather than making assumptions.
6. **CRITICAL: EXTRACT ALL INFORMATION THAT IS PRESENT IN THE REPORT.**
   - Even if information appears in unexpected sections, you must still extract it.
   - Pay special attention to sections like "TECHNICAL QUALITY", "FINDINGS", "CALCIUM SCORE", etc.
   - Look for mentions of vessel conditions throughout the entire report.
   - For calcium scores, gather information about individual vessel scores and total scores.
   - Check for statements about dominance (e.g., "right coronary artery dominance").
7. **Standardize the output formatting as follows:**
   - **CADRADS Score** should only be one of the following: `"0"`, `"1"`, `"2"`, `"3"`, `"4"`, `"5"`, `"N"`. No other values are allowed.
   - **CRITICAL: Use "N" for CADRADS score ONLY when a score is not explicitly provided in the report. DO NOT guess or infer the score based on other information.**
   - **Modifiers** can only be one or more of the following: `"P1"`, `"P2"`, `"P3"`, `"P4"`, `"S"`, `"G"`, `"V"`. If no modifiers are present, use an empty list `[]`.
   - **Only include modifiers that are explicitly mentioned in the report. DO NOT infer modifiers.**
   - **Calcium Score** should be provided as a numerical value only. If a value is given in the report, use that exact value.
   - All **percentages** should follow a consistent format such as `"50-69%"`.
   - If stenosis is present, specify the severity clearly in the `"stenosis_summary"` field, following the format "mild", "moderate", or "severe", accompanied by percentage values.
   - **Ensure all output values are consistent with respect to unit types and word order**. For example, stenosis should be specified uniformly as `"25-49% luminal stenosis"`.

8. **Mapping specific report sections to JSON fields:**
   - For `"left_main"`: Look for LM, Left Main, or mentions of the main trunk of left coronary artery
   - For `"lad"`: Look for LAD, Left Anterior Descending, or anterior interventricular branch
   - For `"lcx"`: Look for LCX, Left Circumflex, circumflex branch
   - For `"rca"`: Look for RCA, Right Coronary Artery
   - For `"calcium_score"`: Look for total calcium score or Agatston score values
   - For `"cardiac_valves"`: Look for descriptions of aortic, mitral, tricuspid, or pulmonic valves
   - For `"pericardium"`: Look for descriptions of the pericardial space/sac/contour
   - For `"dominance"`: Look for statements about coronary dominance pattern (right, left, co-dominant)
   - For `"technical_quality"`: Look for mentions of image quality (excellent, good, fair, poor)
   - For `"impression"`: Look for overall findings or impression sections
9. **Carefully check each section of the report for the features. Do not skip any information that is present.**
   - Thoroughly scan the entire document for relevant information, not just labeled sections.
   - Check alternative spellings and medical terminology variations.
   - When you see a section like "CALCIUM SCORE:" with detailed breakdown, extract both individual vessel scores and the total score.
10. **Ensure the final JSON output is well-formatted and validates correctly.**
11. **If any specific value is unavailable, provide an empty string or `null` instead of additional descriptive content.**
12. **FINAL CHECK: Before submitting, verify that:**
    - You have not hallucinated any values, especially for the CADRADS score and modifiers
    - You have not missed any information that is clearly present in the report
    - You have extracted technical quality, dominance, cardiac valves, and pericardium information if present
    - You have correctly formatted calcium scores as numerical values

Table 11: LLM Prompt for Structured CCTA Report Parsing