NTCIR-18: Proceedings of the 18th NTCIR Conference on Evaluation ofInformation Access Technologies, June 10-13, 2025, Tokyo, Japan
DOI: https://doi.org/10.20736/0002002064

From Divergent LLM Predictions to Reliable Lung Cancer Staging with Ensemble Fusion: CYUT at the NTCIR-18 RadNLP Main Task

Tsz-Yeung Lau
Chaoyang University of Technology
Taichung, Taiwan (R.O.C)
s11327605@gm.cyut.edu.tw

Shih-Hung Wu*
Chaoyang University of Technology
Taichung, Taiwan (R.O.C)
shwu@cyut.edu.tw

Abstract

This study investigates the application of Large Language Models (LLMs) for automated lung cancer staging based on radiology reports, as part of the CYUT team's participation in the NTCIR-18 RadNLP Main Task. Through data analysis, we observed a moderate correlation among the T, N, and M staging classes. Experimental results indicated that jointly prompting LLMs to predict all three classes simultaneously yields improved performance. Additionally, standardizing measurement units to millimeters, rather than centimeters, proved to be a more effective strategy. Based on these findings, we refined our prompting methodology and applied it to both LLMs and reasoning-augmented models, including OpenAI's O-series and DeepSeek-R1. These reasoning-models, enhanced through post-training with Chain-of-Thought (CoT) reasoning, demonstrated superior staging accuracy. As LLMs are generative models, their outputs may vary across different runs, introducing inconsistency in predictions. To mitigate this variability, we adopted an ensemble learning strategy aimed at consolidating divergent LLM outputs into a more stable and reliable lung cancer staging system. Experimental results demonstrate that ensemble methods consistently outperform individual models, enhancing both the robustness and reliability of staging from radiology reports. Our approach achieved second place in the NTCIR-18 RadNLP Main Task (English), underscoring the effectiveness of LLM-based ensemble techniques for TNM classification. The implementation is available at github: anson70242/NTCIR-18-RadNLP-CYUT.

Keywords

Deep Learning, Radiology, NLP, LLMs, Prompt Engineering, Ensemble Learning, Reasoning Models

Team Name

CYUT

Subtasks

Main task (English track)

1 Introduction

The precise and efficient generation of radiology reports is a fundamental aspect of contemporary medical practice, particularly within oncology. However, the increasing volume of medical imaging studies imposes substantial challenges on radiologists, often constraining their ability to produce comprehensive reports within limited time frames[16].

As a result, automating certain components of radiology report analysis has emerged as a focal area of research. Natural language processing (NLP) techniques, particularly those utilizing large language models (LLMs), have shown potential in improving both the efficiency and accuracy of these tasks. Previous studies, including the NTCIR-16 Real-MedNLP [25] and the NTCIR-17 MedNLP-SC Radiology Report TNM Classification Subtask (RR-TNM) [27], as detailed in Table 7, have explored the application of NLP methodologies for radiology report processing. Despite these efforts, findings from the NTCIR-17 RR-TNM task indicate that even SOTA(state-of-the-art) systems, such as GPT-3.5[5] Turbo and GPT-4[17], achieved a maximum joint TNM staging accuracy of only 0.37. This highlights the persistent challenges in applying NLP to medical text classification and underscores the necessity for methodological advancements.

Recent progress in LLM architectures has introduced models such as OpenAI's O-series[18] and DeepSeek-R1[9], which incorporate post-training with Chain-of-Thought (CoT) reasoning data to enhance inferential capabilities. These models hold promise for medical applications, particularly in structured classification tasks. However, despite these advancements, prior research has not systematically assessed the efficacy of these newer models for lung cancer staging based on radiology reports. Additionally, a critical limitation of LLMs is their inherent variability in output predictions across multiple inference runs, which complicates efforts to ensure consistent and reliable classification outcomes. Addressing this issue necessitates methodological approaches that improve prediction stability without incurring substantial computational costs associated with extensive sampling.

Building on prior research on LLM robustness, particularly studies demonstrating the benefits of large-scale sampling for enhancing reasoning accuracy, this study investigates ensemble learning as a viable alternative[4]. While increasing sample size has been shown to improve LLM-based classification performance, practical constraints on computational resources often limit the feasibility of this approach. To address this challenge, we propose an ensemble fusion framework that consolidates predictions from multiple LLMs to enhance both reliability and accuracy in TNM staging. By employing ensemble learning techniques, such as weighted majority voting and XGBoost[6], the proposed framework seeks to mitigate inconsistencies in individual model predictions while leveraging the collective strengths of multiple reasoning models.

In summary, we introduce a novel ensemble-based LLM framework for lung cancer staging that aims to improve both accuracy

*Corresponding author.

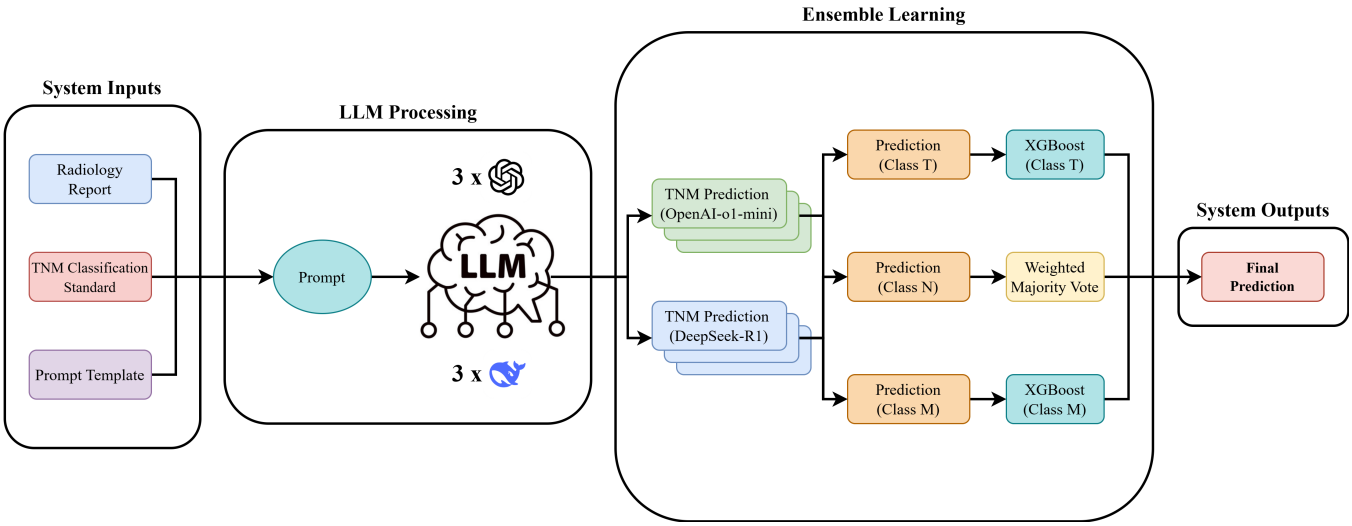

Figure 1: Flow of TNM classification system using LLMs and ensemble learning.

and robustness. The following research questions are addressed in this work:

(1) What is the impact of measurement units (e.g., centimeters versus millimeters) on model performance?
(2) Is it more effective to prompt the LLM to predict staging classes individually or simultaneously?
(3) How do state-of-the-art (SOTA) reasoning models perform in the context of TNM staging?
(4) How to choose the ensemble method based on analysis results during the development stage?

2 Related Work

The utilization of Natural Language Processing (NLP) for medical text analysis has been extensively investigated, with increasing attention directed toward deep learning-based methodologies. Notably, transformer-based architectures, including BERT and its derivatives, as well as large language models (LLMs), have demonstrated notable advancements in medical document classification by effectively capturing contextual representations[3, 21]. Deep learning approaches have significantly enhanced the extraction of clinically relevant information from unstructured medical texts, thereby improving the precision and efficiency of automated analysis. However, the automated TNM staging of radiology reports remains a formidable challenge due to the inherent complexity and ambiguity of medical language. Prior research efforts, such as those undertaken in the NTCIR-16 Real-MedNLP[25] and NTCIR-17 MedNLP-SC Radiology Report TNM Classification Subtask (RR-TNM)[27], have explored deep learning techniques for processing radiology reports. Nonetheless, even state-of-the-art models have achieved only limited accuracy, underscoring the necessity for further methodological advancements to enhance the robustness and reliability of medical text classification.

3 Dataset

3.1 Data Distribution

The datasets used for training and validation in this paper were supplied by the RadNLP organizers[16], comprising a total of 162 samples. Of these, 108 samples are designated for training, while 54 samples are allocated for validation. The dataset contains columns corresponding to the radiology report as well as the T, N, and M classifications. There are several issues in the dataset, including data imbalance, discrepancies between the training and validation sets, and the absence or very low representation of certain classes in the dataset.

Figure 2 illustrates the class distributions. In class T, T4 is the most frequent (28.7% training, 33.3% validation). T0, T1mi, and T1a are absent. T1 is missing in training but constitutes 14.8% of validation. T1b appears in training (8.3%) but not in validation. T1c and T4 are more common in validation, while T2b and T3 are more prevalent in training.

For class N, N0 and N2 dominate (>70%). The training set has 10.1% fewer N0 samples but 6.6% more N1 samples than validation.

In class M, M0 is the majority, with 18.5% more samples in training. M1a is absent in training but represents 16.7% of validation. M1b appears only in training (13%), while M1c is 14.8% lower in training than in validation.

In summary, the dataset presents several challenges across the T, N, and M classifications. These imbalances could negatively affect the performance of traditional methods, such as Artificial Neural Networks (ANNs) and Bidirectional Encoder Representations from Transformers (BERT). However, leveraging pre-trained Large Language Models (LLMs), which are capable of incorporating contextual information like the TNM classification guidelines, may help mitigate the effects of these challenges.

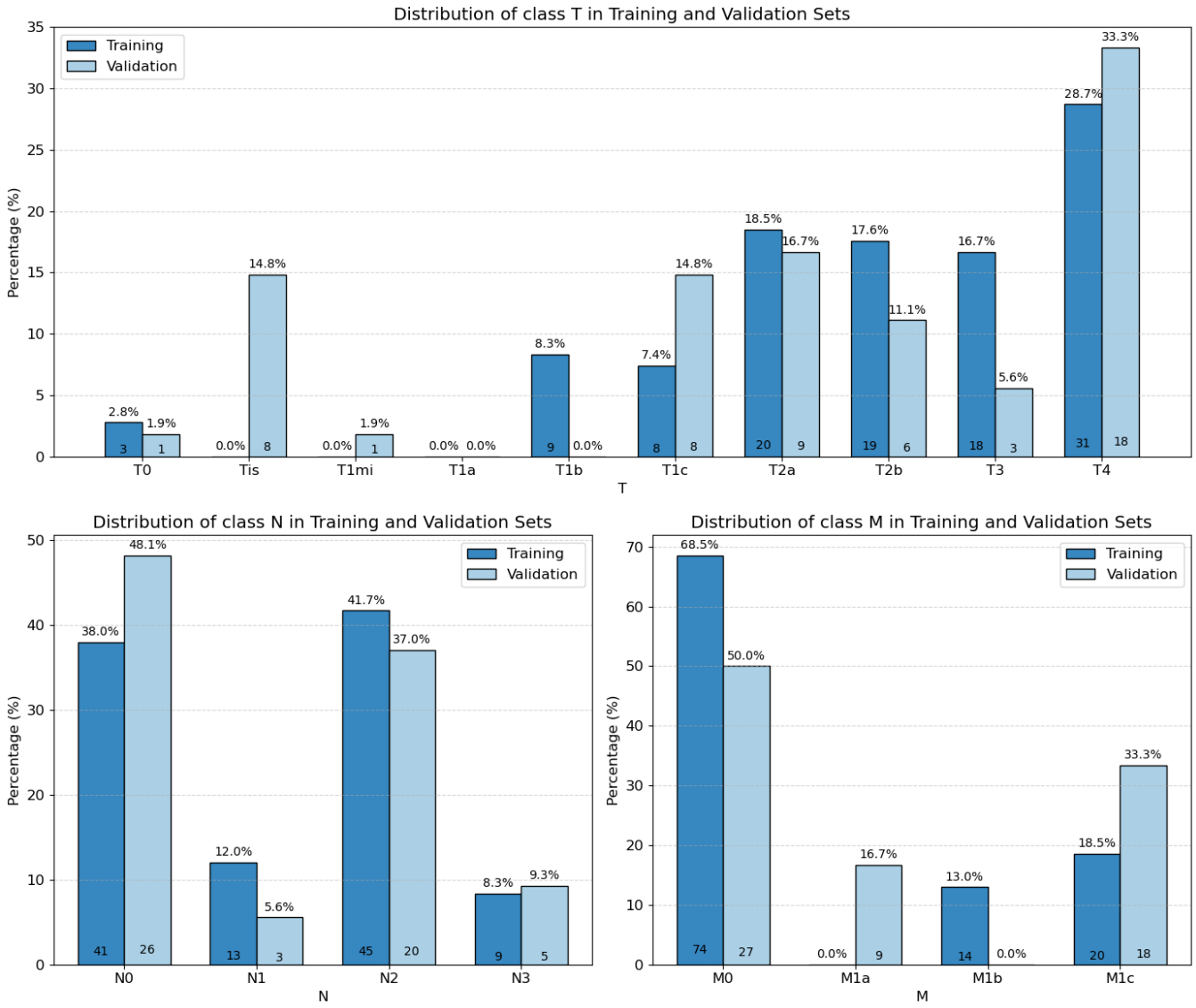

Figure 2: Distribution of class T, class N, and class M across different subsets (Training and Validation).

3.2 Relationships Among TNM Classes

In examining inter-class relationships, Figure 3 presents heatmaps illustrating the co-occurrence patterns among the T, N, and M categories, where the intensity of coloration reflects the frequency of joint occurrences. These visualizations highlight distributional patterns and imply potential interdependencies among the categories, thereby offering a preliminary understanding of their relational structure. Complementarily, Table 1 reports the outcomes of the Chi-square tests, including p-values and Cramér's V coefficients, which collectively assess the statistical significance and strength of associations between the categorical variables. A p-value below the conventional threshold of 0.05 indicates that the observed association is unlikely to have arisen by chance, thus signifying statistical significance. The Cramér's V statistic, ranging from 0 to 1, provides a normalized measure of association strength, with values approaching 1 denoting a strong relationship and those near 0 indicating a negligible or absent association.

Relation	Chi-square Test p-value	Cramér's V
T vs. N	7.0×10^{-9}	0.42
T vs. M	6.6×10^{-8}	0.41
N vs. M	3.6×10^{-8}	0.33

Table 1: Chi-square results and Cramer's V for pairwise comparisons between classes T, N, and M.

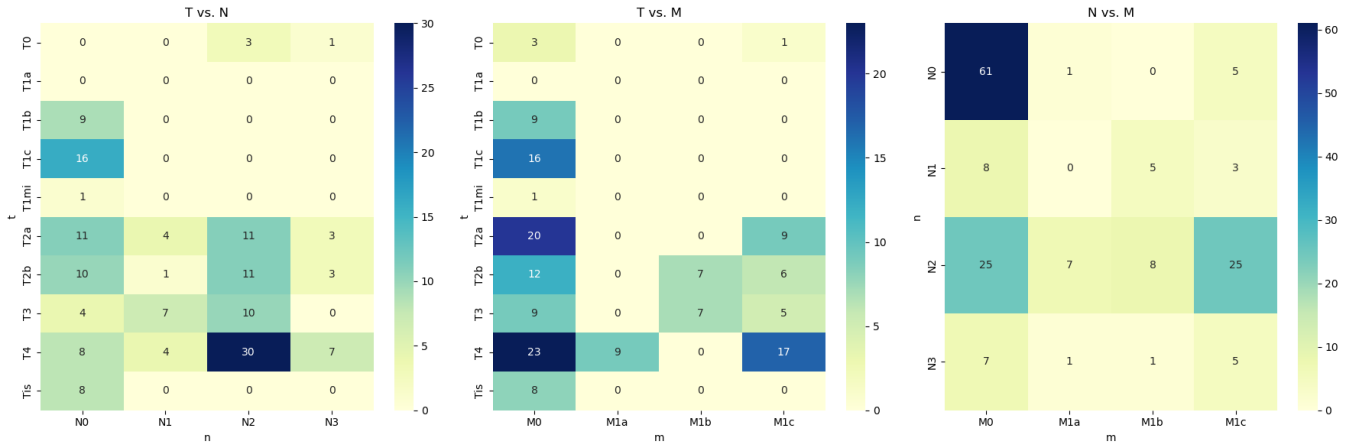

Figure 3: Heatmaps of TNM class relationships.

4 Methods

This study presents a comprehensive framework for automating lung cancer staging through the analysis of radiological reports, as illustrated in Figure 1. The methodology integrates advanced machine learning techniques, including both Large Language Models (LLMs) and traditional natural language processing (NLP) approaches, to convert unstructured clinical narratives into structured staging classifications.

Central to the proposed framework is the application of prompt engineering, wherein carefully crafted input prompts are utilized to optimize LLM performance by directing model attention toward task-specific linguistic and semantic features. In addition, we incorporate reasoning-augmented models—such as OpenAI's o-series and DeepSeek-R1, which utilize Chain-of-Thought (CoT) post-training to enhance multi-step inferential reasoning and clinical decision-making capabilities.

To evaluate the effectiveness of LLMs, we benchmark their performance against RoBERTa [13]. Furthermore, ensemble learning techniques, including Weighted Majority Voting and XGBoost, are employed to synthesize predictions from multiple models, thereby improving overall classification robustness and accuracy. This integrative approach establishes a reliable solution for extracting clinically relevant information from medical text data.

4.1 System Development

Previous research has underscored the challenges LLMs face in performing numerical reasoning, particularly with respect to floating-point operations. To address this limitation, an initial experiment assessed the effect of unit standardization on model accuracy. Converting all measurements to millimeters yielded a statistically significant improvement, indicating that uniform measurement units may enhance the numerical reasoning consistency of language models.

Subsequently, we investigated whether underlying statistical correlations between variables influence the accuracy of multi-label classification. While moderate interdependencies exist among TNM staging variables, concurrent prediction of correlated labels did not inherently improve performance. However, prompting the model to infer interrelated classifications jointly produced a marked increase in accuracy, suggesting that the co-prediction strategy may facilitate more coherent and context-aware inferences.

We also implemented a retrieval-based few-shot learning mechanism, wherein semantically similar prior cases were provided as contextual exemplars. Although this approach showed potential, its efficacy was limited by the dataset's restricted size and diversity, which constrained the representativeness of the retrieved examples. Future applications of this method may benefit from larger and more heterogeneous corpora.

An additional strategy involved instructing the model to generate an intermediate rationale prior to delivering a final classification. This technique aimed to enhance interpretability and guide the model's reasoning process. However, empirical analysis revealed that this prompting approach did not yield measurable improvements in predictive performance.

A detailed summary of experimental results is provided in Table 3. As illustrated by the tokenization examples in Table 2, using the "mm" unit typically converts measurements to integers (e.g., "37 mm") rather than decimals ("3.7 cm"). This likely contributes to improved model performance. Representing sizes as integers provides a consistent format that simplifies numerical comparison for the LLM. Integers are often processed as single tokens (e.g., "37", "103"), potentially making them easier for the model to compare directly than decimal numbers, which may be fragmented into multiple tokens (e.g., "3", ".", "7" or "10", ".", "3") and introduce ambiguity.

4.2 Large Language Models (LLMs)

Large Language Models (LLMs) have become an essential aspect of modern artificial intelligence, significantly impacting computational linguistics and machine learning applications. Proprietary models such as OpenAI's GPT-4o[17], Anthropic's Claude-3.5-sonnet[1], and open-source alternatives including Meta's LLaMA 3[14], Mistral 7B[15], DeepSeek-V3[10], and Qwen 2.5[19], have demonstrated

Report ID	Unit	Text
923073	cm	A mass with a maximum diameter of 3.7 cm in the left upper lobe of the lung.
923073	mm	A mass with a maximum diameter of 37 mm in the left upper lobe of the lung.
4660316	cm	A tumor with a major axis of 10.3 cm is observed in the left lung hilum with obstructive atelectasis of the left lung.
4660316	mm	A tumor with a major axis of 103 mm is observed in the left lung hilum with obstructive atelectasis of the left lung.

Table 2: DeepSeek-R1 Tokenization Examples.

Model	Joint (fine)	T (fine)	N (fine)	M (fine)	Joint (coarse)	T (coarse)	N (coarse)	M (coarse)
Llama3.1-8B-fp16 (mm)	**0.46**	**0.67**	**0.87**	0.85	**0.70**	**0.83**	**0.87**	**0.98**
Llama3.1-8B-fp16 (Few-shot-RAG)	0.41	0.65	0.81	0.74	0.63	0.81	0.81	0.91
Llama3.1-8B-fp16 (cm)	0.35	0.52	0.85	**0.89**	0.46	0.57	0.85	0.98
Llama3.1-8B-fp16 (Reasoning first)	0.30	0.57	0.80	0.83	0.56	0.76	0.80	0.98
Llama3.1-8B-fp16 (Class-By-Class)	0.24	0.39	0.61	0.88	0.27	0.42	0.61	0.91

Table 3: Comparison of Different Approaches with Llama-3 (Sorted by Joint (fine)).

remarkable capabilities in text generation, reasoning, and classification. These systems are widely adopted for automating complex tasks across multiple domains.

To enhance LLM performance, various strategies have been developed, including prompt engineering, post-training adaptation, and test-time computation. Prompt engineering techniques, such as role-playing personas and Chain-of-Thought (CoT) reasoning, can significantly improve inference accuracy on complex tasks. Post-training methods such as Direct Preference Optimization (DPO)[20] and Reinforcement Learning with Human Feedback (RLHF)[7] have also been employed to refine model outputs. At the inference stage, techniques like Beam Search and Monte Carlo Tree Search (MCTS)[26] are used to improve the robustness of predictions through the aggregation of multiple sampled responses. Additionally, test-time computation strategies, such as those explored in Google's Titans study [2], further enhance model performance during deployment.

A notable application of LLMs in this study is the automation of lung cancer staging via the analysis of radiological reports. This task is framed as a text classification problem, in which LLMs are used to extract and categorize relevant medical information from unstructured text, generating structured outputs in standardized formats such as JSON. This approach streamlines the clinical workflow and facilitates efficient downstream processing.

4.3 Reasoning Models

Recent advancements in LLMs have led to the development of models optimized for explicit reasoning tasks. Among these, OpenAI's proprietary o-series (e.g., o1 and o3)[18] and DeepSeek's open-source DeepSeek-R1[9] stand out for their integration of Chain-of-Thought (CoT) data to enhance reasoning capabilities.

DeepSeek-R1 employs a two-phase training procedure. Initially, the model undergoes supervised fine-tuning (SFT), followed by reinforcement learning (RL) to further optimize reasoning performance. To reduce the computational burden typically associated with RL, DeepSeek introduces Group Relative Policy Optimization (GRPO)[8], a novel technique that eliminates the need for a critic model. Instead, GRPO calculates baseline values from group-level scores, enabling more efficient training. The model receives accuracy rewards for correct answers and is further guided by a format reward mechanism that encourages encapsulating its reasoning within `<think>` and `</think>` tags. This promotes the progressive extension of reasoning sequences, improving output precision.

Additionally, the DeepSeek team observed a notable "aha" moment during training: the model initially proposes a solution, then reevaluates and improves upon its reasoning, resulting in a more refined and accurate response. The base model of DeepSeek-R1, DeepSeek-V3[10], employs a Mixture of Experts (MoE)[23] architecture with FP8 mixed-precision training, enhancing computational efficiency without compromising inference quality.

In this study, we leverage these advanced reasoning capabilities of LLMs, particularly for TNM classification in lung cancer staging, aiming to achieve high accuracy and consistent performance despite the inherent stochasticity of LLM outputs. Ensemble-based approaches are further considered to improve the reliability of predictions in structured medical tasks.

4.4 Prompt Engineering

The effectiveness of prompt engineering is critical in enhancing the performance of large language models (LLMs) for domain-specific applications. The formulation of well-structured prompts plays a crucial role in shaping model outputs by directing inference toward accurate and contextually appropriate responses. This study employs systematically designed prompt templates to enhance classification accuracy, integrating semantically rich queries specifically tailored for the analysis of medical reports.

Specifically, our prompt design is finalized through an iterative process of experimentation and refinement. Beginning with a basic structure, we systematically adjust and evaluate its effectiveness based on model performance. Key modifications, such as appending format guidelines, significantly enhance response stability, while explicitly specifying valid classification categories in the instructions helps prevent erroneous predictions. This iterative optimization process enables the prompt to produce consistently reliable outputs with a well-structured and stable format, demonstrating its effectiveness in guiding the model toward accurate and contextually appropriate classifications.

Our prompt consists of four distinct components in the system prompt and three in the user prompt, shown in Table 6. Within the system prompt, we incorporate role-playing[12], which has been demonstrated to enhance LLM capabilities, alongside task-specific instructions, domain-specific knowledge integration (Table 7), and structured output formatting (JSON guidelines). The user prompt encompasses task instructions, radiology report, and output format specifications. To ensure the model precisely understands its function, task instructions and output formatting guidelines are reiterated within the user prompt.

Given this structure, Since OpenAI's O-series models do not support a separate system prompt, we concatenate the system prompt components with the user prompt and pass the combined text as a single user prompt.

4.5 BERT

Bidirectional Encoder Representations from Transformers (BERT)[11] is a transformer-based architecture that has significantly advanced natural language processing by introducing bidirectional contextual learning. RoBERTa[13], an optimized variant of BERT, enhances pretraining methodologies to achieve superior performance. This research employs RoBERTa to compare Large Language Models (LLMs) with conventional natural language processing techniques that preceded the development of LLMs.

4.6 Ensemble Learning

This study employs two ensemble learning techniques: Weighted Majority Voting and eXtreme Gradient Boosting (XGBoost)[6]. The selection of the ensemble method varies based on the classification categories (T, N, and M). The decision-making process for choosing the appropriate ensemble technique involves an in-depth analysis of the predictions generated by Large Language Models (LLMs).

For classes T and M, a substantial proportion of instances exhibit incorrect predictions from five to six individual LLM outputs. Given this higher error rate, XGBoost is applied to enhance predictive accuracy. In contrast, for class N, the majority of LLM predictions are correct in more than half of the instances. Consequently, Weighted Majority Voting is deemed the more suitable approach for this category.

5 Experiments and Result

This section presents the official results of the RadNLP Main Task. The performance of various models, as reported in Table 4, highlights key insights into the effectiveness of different approaches.

The results indicate that reasoning models achieved the highest performance. Specifically, DeepSeek-R1 attained a Joint (fine) accuracy of 0.81, while OpenAI-o1-mini reached 0.78. By applying ensemble learning to these models, combining six predictions in total, the Joint (fine) accuracy improved by 2%, reaching 0.83. This demonstrates the efficacy of ensemble methods in enhancing overall performance by leveraging complementary strengths of multiple models.

An unexpected outcome was observed with GPT-4o, which achieved an accuracy of 0.96 for M (fine), outperforming expectations. Additionally, the LLaMA series exhibited a notable trend: the 8B parameter models performed significantly better than their 70B counterparts, suggesting that model size alone does not guarantee superior performance. This discrepancy may be attributed to architectural differences, optimization strategies, or training data variations.

The distilled model, DeepSeek-R1-70B-Llama-distill, demonstrated a 9% improvement, reinforcing that model distillation remains an effective approach for training smaller models. This finding highlights the potential of knowledge distillation in maintaining high performance while reducing model size.

Furthermore, drawing inspiration from the ML-Promise paper [22], our efforts to improve performance through the application of Retrieval-Augmented Generation (RAG) in the context of few-shot learning did not produce favorable outcomes. Instead of an expected improvement, the Joint (fine) accuracy experienced a substantial decline of 15%. This suggests that the retrieval mechanism employed in this scenario did not effectively support the task, potentially due to domain misalignment or suboptimal retrieval strategies.

We also evaluated additional models on the task. GPT-4o-mini achieved a Joint (fine) accuracy of 0.54, DeepSeek-R1-32B-fp16 reached 0.44, Macro-o1-7B obtained 0.41, Qwen2.5-7B-instruct-fp16 recorded 0.41, and Mistral-nemo-12B-instruct-2407-fp16 achieved 0.35. These results provide further insights into the relative effectiveness of different architectures and model sizes.

As for the test results, shown in Table 5, our best-performing method achieved a Joint (fine) accuracy of 0.60, representing a 23% drop compared to the validation set. This decrease is primarily due to a significant decline in T (fine) accuracy, indicating that our ensemble method for class T was not effective. Further work is needed to refine and improve the approach for this specific class.

Overall, these findings underscore the effectiveness of reasoning-focused models and ensemble learning, while also highlighting the limitations of RAG in this specific context.

Other model's performance will be added after we receive the test set label.

6 Limitation

Upon completing the experiment, we identified certain limitations in our approach. Specifically, the weighting mechanism for the majority vote primarily addresses the same weighting issue rather than functioning as a meaningful weighting system. The enhancement lies in assigning weights to specific seed and temperature settings, thereby ensuring that the assigned weights contribute meaningfully to each prediction.

Model	Joint (fine)	T (fine)	N (fine)	M (fine)	Joint (coarse)	T (coarse)	N (coarse)	M (coarse)
Reasoning Model Ensemble	**0.83**	**0.89**	0.94	**1.00**	**0.93**	**0.98**	**0.94**	**1.00**
DeepSeek-R1	0.81	**0.89**	0.94	0.96	0.89	**0.98**	**0.94**	0.96
OpenAI-o1-mini	0.78	0.81	0.94	0.98	0.87	0.93	**0.94**	0.98
GPT-4o	0.78	0.85	**0.96**	0.93	0.85	0.96	0.96	0.93
Llama3.1-8B-fp16	0.56	0.74	0.87	0.87	0.76	0.91	0.87	0.98
GPT-4o-mini	0.54	0.76	0.93	0.81	0.78	0.91	0.93	0.94
DeepSeek-R1-70B-Llama-distill-q4	0.48	0.69	0.85	0.94	0.76	0.93	0.85	0.98
DeepSeek-R1-32B-fp16	0.44	0.72	0.83	0.80	0.61	0.80	0.83	0.94
Macro-o1-7B[28]	0.41	0.54	0.81	0.93	0.57	0.72	0.81	0.94
Qwen2.5-7B-fp16	0.41	0.57	0.85	0.83	0.70	0.87	0.85	0.96
Llama3.1-70B-q4_0	0.39	0.65	0.87	0.80	0.69	0.91	0.87	0.91
Mistral-nemo-12B-fp16[24]	0.35	0.63	0.85	0.78	0.74	0.94	0.85	0.93
Llama3.3-70B-q4[14]	0.24	0.63	0.65	0.65	0.43	0.83	0.65	0.78
GPT-4o-mini (baseline)	0.24	0.56	0.57	0.80	0.41	0.76	0.57	0.85
RoBERTa	0.24	0.39	0.61	0.88	0.27	0.42	0.61	0.91

Table 4: Validation Accuracies for Various Models(mm) (Sorted by Joint (fine)).

Rank	Joint (fine)	T (fine)	N (fine)	M (fine)	Joint (coarse)	T (coarse)	N (coarse)	M (coarse)
2	**0.60**	**0.69**	**0.94**	**0.93**	**0.63**	**0.70**	**0.94**	**0.94**

Table 5: Results from the Formal Run.

7 Error Analysis

After evaluating the final results, we found that the system does not always make correct predictions. To understand why these mistakes happen, we performed an error analysis.

We used the OpenAI-o1 model, along with knowledge about TNM classification (a system for describing cancer stages), radiology reports, correct labels, and the system's predictions. We then prompted the model to explain why it made classification errors.

For category *N*, the main reasons of misclassification are listed in Table 9.

For the category *T*, as shown in Table 8, the model tends to overestimate how much a tumor has spread when analyzing certain lung abnormalities called *ground-glass lesions*. These lesions are often non-invasive or only slightly invasive, but the model frequently labels them as more advanced cancer stages than they actually are. One common mistake is misclassifying carcinoma *in situ* (Tis), a type of cancer that has not spread, as a more invasive stage (T1b). The main reasons for these errors are:

- **Confusion with Ground-Glass Nodules:** The model misinterprets textual descriptions of pure ground-glass lesions as invasive. *Example: Report ID 241752 – Misclassified a purely ground-glass lesion as T1b despite no invasive component.*
- **Overestimating Tumor Invasion:** The model assigns a higher stage based on lesion size while ignoring minimal invasion criteria. *Example: Report ID 2318717 – Predicted T1b instead of T1mi, overlooking minimal invasion guidelines.*
- **Not Following Medical Guidelines Properly:** The model fails to apply established staging rules correctly. *Example: Report ID 12646171 – Classified as T3 instead of T4, ignoring the rule about tumor spread across lung lobes.*

In summary, these mistakes occur because the model struggles to correctly interpret non-invasive lesions, overestimates invasion, or fails to fully apply TNM classification guidelines.

8 Conclusions and Future Work

8.1 Conclusions

This study presents a robust ensemble framework leveraging Large Language Models (LLMs) for automated lung cancer staging from radiology reports. Our investigation highlights the effectiveness of reasoning-enhanced models such as DeepSeek-R1 and OpenAI's o-series, which outperform conventional baselines in TNM classification. To address the inherent variability of LLM outputs, we proposed an ensemble fusion strategy that aggregates multiple model predictions, significantly improving reliability and accuracy.

Through systematic experimentation, we found that joint prediction of staging classes, standardization of measurement units, and the use of reasoning-augmented models substantially enhance classification performance. The ensemble approach, combining

Weighted Majority Voting and XGBoost based on class-specific characteristics, proved effective in stabilizing predictions.

Our system achieved second place in the NTCIR-18 RadNLP Main Task (English), validating the practical value of ensemble-based LLM strategies for medical NLP applications. The findings underscore the potential of combining diverse reasoning models to deliver more consistent and trustworthy AI-assisted cancer staging tools.

8.2 Future Work

Future research directions should focus on addressing identified limitations and enhancing model robustness. One promising avenue is the incorporation of more advanced models, such as OpenAI's o3 or ModernBERT, which may offer improved contextual understanding and predictive accuracy. Additionally, as the availability of annotated radiology reports increases, fine-tuning LLMs using techniques like supervised fine-tuning (SFT) or test-time computing could further improve performance.

Another important direction is the exploration of Retrieval-Augmented Generation (RAG) methods to enhance few-shot learning. With a larger dataset, RAG could enable LLMs to retrieve and leverage similar cases, thereby improving classification consistency.

Lastly, integrating multimodal learning approaches that incorporate both textual and imaging data holds promise for further improving automated TNM staging. Future studies should explore synergistic methodologies that combine NLP-driven text classification with image-based tumor analysis, facilitating a more comprehensive AI-driven radiology workflow.

Acknowledgments

This study was supported by the National Science and Technology Council under the grant number NSTC 113-2221-E-324-009.

References

[1] Anthropic. 2024. Claude 3.5 Sonnet Model Card Addendum. arXiv:2407.21787 [cs.LG] https://arxiv.org/abs/2407.21787

[2] Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. 2024. Titans: Learning to Memorize at Test Time. arXiv:2501.00663 [cs.LG] https://arxiv.org/abs/2501.00663

[3] Pavel Blinov, Manvel Avetisian, Vladimir Kokh, Dmitry Umerenkov, and Alexander Tuzhilin. 2020. *Predicting Clinical Diagnosis from Patients Electronic Health Records Using BERT-Based Neural Networks.* Springer International Publishing, 111–121. https://doi.org/10.1007/978-3-030-59137-3_11

[4] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large Language Monkeys: Scaling Inference Compute with Repeated Sampling. arXiv:2407.21787 [cs.LG] https://arxiv.org/abs/2407.21787

[5] Tom B. Brown et al. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

[6] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[7] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. arXiv:1706.03741 [stat.ML] https://arxiv.org/abs/1706.03741

[8] DeepSeek-AI. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300 [cs.CL] https://arxiv.org/abs/2402.03300

[9] DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] https://arxiv.org/abs/2501.12948

[10] DeepSeek-AI. 2025. DeepSeek-V3 Technical Report. arXiv:2412.19437 [cs.CL] https://arxiv.org/abs/2412.19437

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[12] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better Zero-Shot Reasoning with Role-Play Prompting. arXiv:2308.07702 [cs.CL] https://arxiv.org/abs/2308.07702

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL] https://arxiv.org/abs/1907.11692

[14] Meta-AI. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[15] Mistral-AI. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[16] Yuta Nakamura, Michael Krauthammer, Tomohiro Kikuchi, Peitao Han, Shouhei Hanaoka, Koji Fujimoto, and et al. 2025. NTCIR-18 RadNLP 2024 Overview: Dataset and Solutions for Automated Lung Cancer Staging.

[17] OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[18] OpenAI. 2024. OpenAI o1 System Card. https://cdn.openai.com/o1-system-card-20241205.pdf Accessed: 2025-02-24.

[19] QwenTeam. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[20] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290 [cs.LG] https://arxiv.org/abs/2305.18290

[21] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine* 4, 1 (May 2021). https://doi.org/10.1038/s41746-021-00455-y

[22] Yohei Seki, Hakusen Shu, Anaïs Lhuissier, Hanwool Lee, Juyeon Kang, Min-Yuh Day, and Chung-Chi Chen. 2024. ML-Promise: A Multilingual Dataset for Corporate Promise Verification. arXiv:2411.04473 [cs.CL] https://arxiv.org/abs/2411.04473

[23] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. arXiv:1701.06538 [cs.LG] https://arxiv.org/abs/1701.06538

[24] Sharath T. Sreenivas, Saurav Muralidharan, Raviraj Joshi, Marcin Chochowski, Ameya S. Mahabaleshwarkar, et al. 2024. LLM Pruning and Distillation in Practice: The Minitron Approach. arXiv:2408.11796 [cs.CL] https://arxiv.org/abs/2408.11796

[25] Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. 2022. RealMedNLP: Overview of REAL document-based MEDical Natural Language Processing Task SUBTASKS.

[26] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601 [cs.CL] https://arxiv.org/abs/2305.10601

[27] Yuta Nakamura, Shouhei Hanaoka, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2023. NTCIR-17 MedNLP-SC Radiology Report Subtask Overview: Dataset and Solutions for Automated Lung Cancer Staging. , none pages.

[28] Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards Open Reasoning Models for Open-Ended Solutions. arXiv:2411.14405 [cs.CL] https://arxiv.org/abs/2411.14405

Appendices A–B

A Prompt Template and TNM Classification Guidelines

B Misclassified cases

Type of Prompt	Element	Prompt
System prompt	Role playing	You are a seasoned oncologist specializing in cancer classification.
	Instruction	Your task is to determine the TNM classification based solely on the provided radiology report. —
	Domain knowledge	TNM Classification System Overview: - T (Tumor): Size and extent of the primary tumor. - N (Node): Whether the cancer has spread to nearby lymph nodes. - M (Metastasis): Whether the cancer has spread to distant parts of the body. {TNM Classification Guideline} —
	Instruction	Instructions: - Use only the information from the radiology report. - Provide your classification strictly in the specified JSON format without any additional text. - Ensure each classification is specific to the categories provided. - Include an additional key "Reason" in the JSON output explaining the rationale behind your TNM classification. —
	Output control	Classification Options: - T: {T0, Tis, T1mi, T1a, T1b, T1c, T2a, T2b, T3, T4} - Note: "T1" should be broken down into T1mi, T1a, T1b, or T1c. - Note: "T2" should be specified as either T2a or T2b. - Note: Use T0 if the tumor does not fit into Tis to T4. - N: {N0, N1, N2, N3} - M: {M0, M1a, M1b, M1c} - Note: "M1" must be specified as M1a, M1b, or M1c; there is no "M1" category. —

	Format control	Example Output:
		{
		"T": "T2a",
		"N": "N1",
		"M": "M0",
		"Reason": "Your explanation for the TNM classification."
		}

User prompt	Instruction	Please derive the TNM classification based on the following radiology report. Use only the information provided and ensure that the classification follows the format below without any additional text. —
	Radiology Report	Radiology Report: {radiology report} —

	Format control	Classification Format (choose from the specified options):
		{
		"T": "T?",
		"N": "N?",
		"M": "M?",
		"Reason": "Your explanation for the TNM classification."
		}

	Output control	Classification Options: - T: {T0, Tis, T1mi, T1a, T1b, T1c, T2a, T2b, T3, T4} - Note: There is no "T1" or "T2" without subcategories. - N: {N0, N1, N2, N3} - M: {M0, M1a, M1b, M1c} - Note: There is no "M1" without subcategories.

Table 6: Prompt Template.

TNM Classification Guidelines

—
T: primary tumor
- T0: *no evidence of a primary tumor or primary tumor cannot be assessed*
- Tis: *carcinoma in situ - tumor measuring 30 mm or less and has no invasive component at histopathology*
- T1: *tumor measuring 30 mm or less in greatest dimension surrounded by lung or visceral pleura without bronchoscopic evidence of invasion more proximal than the lobar bronchus (i.e. not in the main bronchus)
  - T1mi: minimally invasive adenocarcinoma
  - tumor has an invasive component measuring 5 mm or less at histopathology and the overall tumor size not exceed 30 mm.
  - T1a: tumor less than or equal to 10 mm in greatest dimension
  - T1b: tumor greater than 10 mm but less than or equal to 20 mm in greatest dimension
  - T1c: tumor greater than 20 mm but less than or equal to 30 mm in greatest dimension
- T2: tumor greater than 30 mm but less than or equal to 50 mm or tumor with any of the following features:
  involves the main bronchus regardless of distance from the carina but without the involvement of the carina
  invades visceral pleura
  * associated with atelectasis or obstructive pneumonitis that extends to the hilar region (involving part or all of the lung)*
  - T2a: tumor greater than 30 mm but less than or equal to 40 mm in greatest dimension
  - T2b: tumor greater than 40 mm but less than or equal to 50 mm in greatest dimension
- T3: tumor greater than 50 mm but less than or equal to 70 mm in greatest dimension or associated with separate tumor nodule(s) in the same lobe as the primary tumor or directly invades any of the following structures:
  chest wall (including the parietal pleura and superior sulcus)
  phrenic nerve
  parietal pericardium
- T4: *tumor greater than 70 mm in greatest dimension* or
  associated with separate tumor nodule(s) in a different ipsilateral lobe than that of the primary tumor or
  *invades any of the following structures:
  diaphragm, mediastinum, heart, great vessels, trachea, recurrent laryngeal nerve, sophagus, vertebral body, carina*
—
N: *regional lymph node involvement*
- N0: no regional lymph node metastasis or regional lymph nodes cannot be assessed or regional lymph nodes cannot be assessed
- N1: metastasis in ipsilateral peribronchial and/or ipsilateral hilar lymph nodes and intrapulmonary nodes, including involvement by direct extension
- N2: metastasis in ipsilateral mediastinal and/or subcarinal lymph node(s)
- N3: metastasis in contralateral mediastinal, contralateral hilar, ipsilateral or contralateral scalene, or supraclavicular lymph node(s)
—
M: *distant metastasis*
- M0: no distant metastasis
- M1: distant metastasis present
  - M1a: separate tumor nodule(s) in a contralateral lobe; tumor with pleural or pericardial nodule(s) or malignant pleural or pericardial effusions
  - M1b: single extrathoracic metastasis, involving a single organ or a single distant (nonregional) node
  - M1c: multiple extrathoracic metastases in one or more organs
—

Table 7: TNM Classification Guidelines.

Report ID	Label (T)	Prediction (T)	Reason
241752	Tis	T1b	The prediction is incorrect because the lesion is purely ground-glass and lacks any solid or invasive component. According to the guidelines, a noninvasive lesion less or equal to 30 mm is classified as Tis (carcinoma in situ), not T1b.
2318717	T1mi	T1b	Because the tumor is described as a minimally invasive adenocarcinoma (invasive component less or equal to 5 mm) and is only 15 mm in total size, it fits T1mi, not T1b. The predicted classification overlooks the minimal invasion criterion and classifies based solely on size.
4592263	Tis	T1b	The nodule is purely ground-glass and shows no solid (invasive) component, which corresponds to an in-situ lesion (Tis). T1b would require at least some evidence of invasive disease, so classifying the tumor as T1b rather than Tis is incorrect in this context.
12646171	T4	T3	The prediction incorrectly classified the primary tumor as T3 despite evidence that it extends from the right upper lobe into the right middle lobe. Having tumor nodules in different lobes of the same lung automatically makes it T4 (even if there is chest wall invasion). Therefore, the correct classification is T4, not T3.
16191878	Tis	T1b	Because the nodule is purely ground-glass (no known invasive component) and measures 15 mm, it fits the definition of carcinoma in situ (Tis). T1b implies an invasive tumor larger than 10 mm, which contradicts the "pure" GGN (noninvasive) finding. Hence, T1b is incorrect; Tis, N0, M0 is more appropriate.
16572985	Tis	T1b	Because the lesion is a pure GGN with no identified invasive component on imaging and is 15 mm (less or equal to 30 mm), it fits Tis (carcinoma in situ) rather than T1b. T1b would require an invasive component in a lesion >10 mm. Here, no invasion was demonstrated, so labeling it T1b is incorrect.

Table 8: Misclassified cases for class T.

Report ID	Label (N)	Prediction (N)	Reason
4734929	N0	N2	Because "slightly conglomerated" mediastinal lymph nodes are not definitively malignant on imaging alone, there is no confirmed evidence of N2 disease. Hence, the correct classification remains N0, not N2.
10868892	N2	N3	Because the contralateral mediastinal nodes are only noted as "suspect" (not confirmed) metastases, there is insufficient evidence to assign N3. The radiology findings confirm ipsilateral mediastinal nodal involvement (N2) rather than definitive contralateral involvement (N3).
11566958	N0	N1	Because the small peripheral nodules do not necessarily indicate hilar or peribronchial lymph node metastasis (N1), there is no confirmed nodal involvement on imaging or clinical data. Therefore, labeling them as N1 is incorrect, and the correct classification remains T2aN0M0.

Table 9: Misclassified cases for class N.