UOM at the NTCIR-18 RadNLP Task

Wuraola Oyewusi[1], Eliana Vasquez Osorio[1,2], Goran Nenadic[1], and Gareth Price[1,2]

[1]The University of Manchester, Manchester, UK
[2]The Christie NHS Foundation Trust, Manchester, UK

Background/Aims

- Lung cancer staging is crucial for guiding treatment decisions and predicting patient outcomes.
- Accurate staging is critical, as errors can lead to inappropriate treatment plans, impacting survival and quality of care.
- The RadNLP 2024 shared task at NTCIR-18 addresses this by encouraging the development of NLP techniques for automatically classifying lung cancer stages (T: Tumour size, N: Node involvement, M: Metastasis) from radiology reports.[1]
- Key challenges include the variability of clinical language, limited dataset size, and significant class imbalance across TNM stages. These factors hinder effective model generalization and fair evaluation across all classes.
- This work describes the UoM team's approach, tackling limited data and class imbalance through data augmentation and stratification.
- The focus is on improving generalization and ensuring balanced evaluation across all TNM stages.

Methods

Our methodology (Figure 1) involved dataset preparation, data augmentation, and model selection/training.

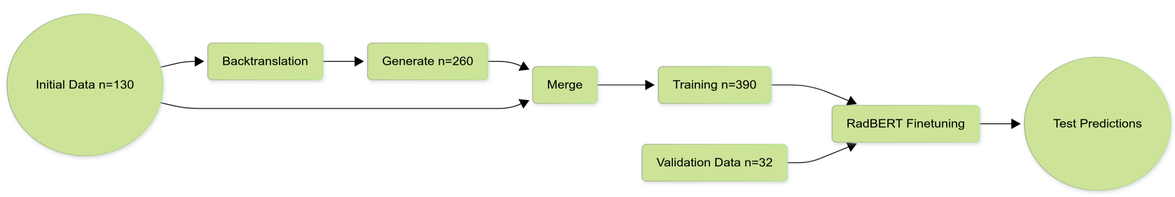

Figure 1: Process Flow for our method for T, N, M prediction

Dataset : The task dataset comprised 162 radiology reports (108 training, 54 validation) with TNM labels. Initial analysis revealed significant class imbalance, and some labels absent in either training or validation splits. Table 1 shows the original data distribution for the 3 classes

	Train	Valid		Train	Valid		Train	Valid
T			N			M		
T0	3	1	N0	41	26	M10	74	27
Tis	0	8	N1	13	3	M1a	0	9
T1mi	0	1	N2	45	20	M1b	14	0
T1b	9	0	N3	9	5	M1c	20	18
T1c	8	8						
T2a	20	9						
T2b	19	6						
T3	18	3						
T4	31	18						

Table 1:Original Data Distribution for T, N, and M Classes

To mitigate imbalance, we applied stratification, selecting a 20% level based on the T class, as it provided the best balance for model performance and validation set size. Stratification by T also improved N and M class distribution.
Table 2 shows the data distribution for T, N, and M classes after a 20% Stratification using T staging

	Train	Valid		Train	Valid		Train	Valid
T			N			M		
T0	3	1	N0	51	16	M10	78	23
Tis	6	2	N1	14	2	M1a	7	2
T1mi	1	0	N2	53	12	M1b	13	1
T1b	7	2	N3	12	2	M1c	32	6
T1c	13	3						
T2a	24	5						
T2b	20	5						
T3	17	4						
T4	39	10						

Table 2: Data Distribution for T, N, and M classes after a 20% Stratification using T staging

While stratification of the original validation set could introduce data leakage, ensuring representation of all classes in both sets was deemed more critical for our small, imbalanced dataset

Data Augmentation : To address the small dataset, we used back-translation (English -> French/German -> English) to increase training data from 130 to 390 instances. This technique introduces textual variation while preserving core meaning and ensuring TNM stage information remained unchanged.
The aim was to improve model generalizability and reduce overfitting

Model Selection and Training : We utilized RadBERT, a transformer model pre-trained on 4,056,227 radiology reports. Its pre-training on radiology reports makes it well-suited for this task. The model is available via Hugging Face. We fine-tuned three separate RadBERT models for T, N, and M stages on the augmented training data using the Transformers library in Python

Experiments and Results

Performance was evaluated using fine-grained accuracy (exact TNM stage match, e.g., T2a vs. T2) and coarse-grained accuracy (broader categories, e.g., T2).
Joint accuracy: proportion of reports with all three TNM stages predicted correctly.

	Validation(Bespoke) (No Augmentation	Validation (Bespoke) (K-Fold =5)	Validation(Task)	Test(Task
Joint Accuracy(Fine)	-	-	**0.9630**	**0.1235**
T Accuracy(Fine)	**0.3939**	**0.9405**	**1.0000**	**0.3333**
N Accuracy(Fine)	**0.6061**	**0.9881**	**0.9815**	**0.5926**
M Accuracy(Fine)	**0.8788**	**1.0000**	**0.9815**	**0.6914**
Joint Accuracy(Coarse)	-	-	**0.9630**	**0.2593**
T Accuracy(Coarse)	**0.4242**	**0.9762**	**1.0000**	**0.4444**
N Accuracy(Coarse)	**0.6061**	**0.9881**	**0.9815**	**0.5926**
M Accuracy(Coarse)	**0.9091**	**1.0000**	**0.9815**	**0.7901**

5-Fold Cross-Validation provided a more robust performance estimate on augmented data, reducing variance but there was a substantial performance drop from validation to test set, indicating poor generalization.

Conclusion

- We explored automated classification of lung cancer TNM stages using RadBERT, focusing on addressing data scarcity and class imbalance via stratified sampling and data augmentation.
- Back-translation improved data diversity and validation accuracy; 5-fold cross-validation provided more reliable performance estimates on augmented data.
- Despite these improvements, the model showed poor generalization to the test set (joint accuracy: 96.3% on validation vs. 12.35% on test).
- This highlights challenges in building models that generalize well with domain-specific language variability and limited data, suggesting overfitting.

References

[1] Yuta Nakamura, Koji Fujimoto, Jonas Kluckert, Michael Krauthammer, Jun Kan zawa, Akira Katayama, Tomohiro Kikuchi, Ryo Kurokawa, Wataru Gonoi, Peitao Han, Kiyoto Hashimoto, Yuki Tashiro, Shouhei Hanaoka, Shuntaro Yada, and Eiji Aramaki. 2025. NTCIR-18 RadNLP 2024 Overview: Dataset and Solutions for Automated Lung Cancer Staging. In Proceedings of the NTCIR-18 Conference.