NTCIR-18: Proceedings of the 18th NTCIR Conference on Evaluation ofInformation Access Technologies, June 10-13, 2025, Tokyo, Japan
DOI: https://doi.org/10.20736/0002002078

TUSNLP at the NTCIR-18 RadNLP Task: Explainable Classification Approach by Domain Knowledge-Based Bag-of-Words

Tomoki Terada
Tokyo University of Science, Japan
8722130@ed.tus.ac.jp

Rei Noguchi*
Tokyo University of Science, Japan
rnoguchi@rs.tus.ac.jp

ABSTRACT

We developed highly interpretable classification models of lung cancer stage using Bag-of-Words representations that consist of predefined key terms based on domain knowledge. These models had high medical validity and provided new clinical insights. This study demonstrates the effectiveness of domain knowledge in improving model accuracy and the usefulness of model interpretability in the medical field.

KEYWORDS

Bag-of-Words, explainability, interpretability, domain knowledge, feature selection, CatBoost, ensemble learning

TEAM NAME

TUSNLP

SUBTASKS

Main task (Japanese track)

1 INTRODUCTION

A radiology report contains the advanced findings of a radiologist. In diagnostic imaging, an attending physician usually orders imaging, and a radiologist first checks the images and describes his or her findings in a radiology report. The attending physician then makes a definitive diagnosis, referring to the images and the radiology report. This means that the author of a radiology report and the final decision maker are different, which can often be factors such as oversight of a radiology report [1].

In the busy clinical field, written descriptions can easily lead to oversights, and it is extremely valuable to extract important findings and critical information leading to a diagnosis from reading reports. In this study, we aim to develop models that automatically determine the stage (degree of progression) of lung cancer from radiology reports and to build a highly interpretable model that is suitable for use in the medical field.

2 RELATED WORK

We previously developed a method for extracting a case matrix, in which each record was unique for each case and the presence or absence of each symptom was stored in separate columns, from discharge summaries [2]. In addition, we applied this case matrix approach in the Real-MedNLP Task of NTCIR-16 to identify similar cases from radiology reports [3].

3 METHODS

3.1 Dataset and overview of our system

A Japanese learning dataset for this study was provided by RadNLP 2024 shared task organizers [4]. This dataset consists of Japanese radiology reports that contain textual information on lung cancer stages based on the TNM classification.

In this study, we developed separate classification models for the T, N, and M categories as models for determining the stage (degree of progression) of lung cancer from radiological reports. In addition, since tumor size is important in determining the stage of the T category, we also developed a model to estimate the major size of the tumor, and utilized the results of the estimation in learning stage prediction.

3.2 Data processing

3.2.1 Extraction of medical terms and calculation of word frequency

As a common preprocessing step for the T, N, and M categories, pre-specified medical terms such as symptoms were first extracted from radiology reports, and the frequency of each term was calculated for each report. This word frequency table was in the form of a kind of “Bag-of-Words” (BoW) and was used for explanatory variables in model development described below.

Three frequency patterns were obtained by applying the same calculation to three sets of text data with different preprocessing. The first word frequency was calculated from the original radiology reports (frequency1). The second word frequency was calculated

* Corresponding author.

from text from which sentences containing pre-specified negative expressions were excluded (frequency2). Note that if they were described in a concession clause (i.e., the expressions were followed by a main clause with contrasting content), they were left in place without exclusion. The third word frequency was calculated from sentences containing "lymph," which is the word closely related to the classification of lung cancer, in addition to excluding negative expressions (frequency3).

3.2.2 Preprocessing for M category

For the M category, we defined key terms related to M1a and M1c based on the staging criteria of the Japan Lung Cancer Society (JLCS) [5]. Since M1c is a stage in which "multiple distant metastases to one or more organs outside the thoracic cavity are observed," we defined the names of organs other than the lungs as key terms for M1c and calculated their frequency in the text data before and after exclusion of negative expressions. On the other hand, since M1a includes the findings of "paraneoplastic nodules in the contralateral lung, pleural or pericardial nodules, malignant pleural effusion (ipsilateral and contralateral), and malignant pericardial effusion," we defined these related words as key terms for M1a and calculated their frequency for text data before and after exclusion of negative expressions as well.

3.2.3 Preprocessing for N category

For the N category, we defined key terms related to N1, N2, and N3 based on the staging criteria of JLCS as well. Since N1 is defined as "metastasis to the ipsilateral peribronchial and/or ipsilateral pulmonary hilum or intrapulmonary lymph nodes to include direct invasion of the primary tumor," we defined the peribronchial and hilum sites as key terms for N1.

Similarly, N2 is defined as "metastasis to lymph nodes in the ipsilateral mediastinum and/or sub-tracheal bifurcation," and we defined mediastinal and sub-tracheal sites as key terms for N2.

Lastly, N3 is defined as "metastasis to the contralateral mediastinum, contralateral pulmonary hilum, ipsilateral or contralateral anterior scalene muscle, or supraclavicular fossa lymph nodes," and therefore, we defined sites related to the contralateral mediastinum, contralateral pulmonary hilum, anterior scalene muscle, and supraclavicular fossa as key terms for N3.

After these definitions, we calculated the frequencies of the key terms by category from sentences containing "lymph."

3.2.4 Preprocessing for T category

For the T category, we extracted the size of the tumor or mass from the text data. We extracted the maximum value when multiple values were present in the text data.

3.3 Model development

We developed prediction models for M, T, and N classifications in lung cancer cases by machine learning and developed a regression model to predict the maximum tumor diameter. Random Forest was used to predict M and T classifications, and ensemble learning combining Random Forest and LightGBM (Light Gradient Boosting Machine) was employed to predict N classification. CatBoost Regressor [6] was used to predict the maximum tumor diameter, and the model was optimized and evaluated.

3.3.1 Classification model in M and T categories

3.3.1.1 Feature selection

First, unnecessary variables, such as identifiers, were excluded from the training data, and variables used for modeling were selected. Then, the importance of the features was calculated using Random Forest, and only variables whose importance exceeded a threshold value (0.001) were extracted. This feature selection reduced computational cost and improved the interpretability of models while maintaining their prediction accuracy.

3.3.1.2 Model training and hyperparameter optimization

Random Forest classifier was trained using the selected important features. In this study, class_weight="balanced" was set to address class imbalances, assigning weights according to the number of samples in each class.

Hyperparameters were optimized using grid search (GridSearchCV). The hyperparameters to be searched were as follows:

- Maximum depth of decision trees (max_depth): None, 10, 20, 30
- Number of decision trees (n_estimators): 100, 200, 300
- Minimum number of samples for node splits (min_samples_split): 2, 5, 10
- Minimum number of leaf nodes (min_samples_leaf): 1, 2, 4

K-fold cross-validation (k=3) was performed, and the best model was selected by optimizing the hyperparameters based on the weighted F1-score indicator.

3.3.2 Classification model in N category

3.3.2.1 Feature selection

Similarly, for the development of the N classification model, unnecessary variables, such as identifiers, were excluded from the training data, and essential features were selected based on feature importance calculated by Random Forest. Note that, unlike the M and T classifications, the threshold value of feature selection criteria for the N classification was set at 0.005.

3.3.2.2 Data standardization and imbalanced data handling

After feature selection, the data was standardized to enable consistent learning of Random Forest and LightGBM. In addition, to alleviate the class imbalance in the data, the classes were balanced by combining samples from a small number of classes with SMOTE (Synthetic Minority Over-sampling Technique).

3.3.2.3 Model training and hyperparameter optimization

The N classification model was optimized by randomized search (RandomizedSearchCV). The following hyperparameters were explored：

- Number of decision trees (n_estimators)
- Maximum depth (max_depth)
- Minimum number of samples for node splits (min_samples_split)
- Minimum number of leaf node samples (min_samples_leaf)
- Feature selection method (max_features)

K-fold cross-validation (k=3) was performed, and the best model was selected by optimizing the hyperparameters based on the weighted F1-score indicator.

Ensemble learning by soft voting was performed to integrate the optimized Random Forest and LightGBM models. We aimed to develop a more robust classification model by averaging the prediction probabilities of each model with the voting classifier.

3.3.3 Maximum tumor diameter prediction model

3.3.3.1 Data processing

For the prediction model of maximum tumor diameter, only cases with max_mm other than zero were used as the training data excluding those with max_mm = 0. In addition, unnecessary variables, such as identifiers, were excluded, and the data was split into training and test data at a ratio of 80% and 20%.

3.3.3.2 Model training and hyperparameter optimization

CatBoost Regressor was used for the maximum tumor diameter prediction model, and the following hyperparameters were optimized by grid search (GridSearchCV):

- Number of iterations (iterations): 500, 1000
- Learning rate (learning_rate): 0.01, 0.05, 0.1
- Tree depth (depth): 4, 6, 8

K-fold cross-validation (k=3) was performed, and the hyperparameters were optimized based on RMSE (Root Mean Squared Error).

3.3.3.3 Handling of cases with a tumor size of 0

The optimized CatBoost model was applied to the cases with max_mm = 0, which was not included in the training data, and tumor sizes of the cases were estimated. This enabled us to provide appropriate estimates even for cases with unrecorded tumor sizes.

3.4 Model evaluation

3.4.1 Train-test data splitting

In this study, the data was split into training and test data at 80% and 20% ratios while maintaining the class distribution by stratified extraction. This allowed the models to be evaluated on the test data with the same distribution as the training data, leading to the models with better generalization performance.

3.4.2 Model evaluation metrics

The classification models were evaluated based on general evaluation metrics such as precision, recall, F1-score and weighted F1-score. Weighted F1-score is often used in imbalanced data, enabling us to improve accuracy while adequately evaluating the impact of classes with a small number.

3.4.3 Feature importance visualization

As described in Section 3.3.1.1 and 3.3.2.1, feature importance was calculated by Random Forest, and only variables that exceeded the specified threshold (0.01 for M and T classifications and 0.05 for N classification) were selected.

In addition, AI in the medical field requires the explainability of models. The machine learning methods used in this study, such as Random Forest and LightGBM, can visualize the importance of each feature. This visualization makes it easier to understand which features are important to the model and contributes to improving interpretability.

4 EXPERIMENTS

4.1 Training and validation results

Table 1 shows training and validation results of Joint, T, N, and M accuracies in terms of “fine” and “coarse,” which are the specified evaluation metrics in RadNLP Task. The T, N, and M classification accuracies (fine) in the train data were 0.9630, 0.9537, and 0.9815, respectively, and the Joint accuracy (fine) was 0.9074. Similarly, the accuracy (fine) of the T, N, and M classifications in the validation data were 0.9630, 0.9444, and 0.9815, respectively, and the Joint precision (fine) was 0.9074. Especially for the N classification, there was a tendency to misclassify N0 as N2.

Table 1 Training, validation, and test (formal run) results

	Fine				Coarse			
	Joint accuracy	T accuracy	N accuracy	M accuracy	Joint accuracy	T accuracy	N accuracy	M accuracy
Train	0.9074	0.9630	0.9537	0.9815	0.9074	0.9630	0.9537	0.9815
Validation	0.9074	0.9630	0.9444	0.9815	0.9259	0.9815	0.9444	0.9815
Test (formal run)	0.2176	0.3519	0.8287	0.7963	0.3796	0.5000	0.8287	0.8611

4.2 Feature importance and model validity

4.2.1 T classification

In evaluating the importance of the features contributing to the prediction of T classification, the feature with the highest contribution was "max_mm" (0.0438), as shown in Table 2. max_mm is a feature indicating the maximum diameter of the tumor, which is a medically valid result since one of the criteria for T classification is the size of the tumor.

The next most important features were "**左**_frequency1" (0.0216), "**左**_frequency2" (0.0210), and "**縦 隔**_frequency2" (0.0206), which mean "left" and "mediastinum" in English, respectively. These results suggest that tumor localization may influence T classification. "**浸潤**_frequency1" (0.0192) and "**浸潤**_frequency2" (0.0185), which mean "infiltration" in English, also showed high importance, consistent with T classification being based on local extension of the tumor.

On the other hand, features such as "**転 移**_frequency2" (0.0114), "**リンパ節**_frequency2" (0.0098), and "N0" (0.0097), which mean "metastasis," "lymph node" and N0 stage in English, respectively, also showed a certain level of importance. Although T classification is essentially a measure of local tumor extension and is independent of N classification (lymph node metastasis) and M classification (distant metastasis), clinically high T classification cases are often accompanied by lymph node metastasis, which may have influenced the model learning.

Table 2 Feature importances in T classification

	Feature	Importance
1	max_mm	0.043783
2	左_frequency1	0.021610
3	左_frequency2	0.021019
4	縦隔_frequency2	0.020570
5	浸潤_frequency1	0.019226
6	浸潤_frequency2	0.018517
7	転移_frequency2	0.011432
8	リンパ節_frequency2	0.009752
9	N0	0.009691
10	腫瘤_frequency1	0.009628

4.2.2 N classification

Table 3 illustrates that in the N classification prediction, "**リンパ節**_frequency3" (0.0470) showed the highest contribution, followed by "**リ ン パ 節**_frequency2" (0.0380), "**リ ン パ 節**_frequency2" (0.0364), and "**リンパ節**_frequency3" (0. 0310). All these features indicated the frequency of words associated with lymph node metastasis, a result consistent with the N classification being determined based on the presence or absence of mediastinal lymph node metastasis.

"**腫大**_frequency3" (0.0244) and "**腫大**_frequency2" (0.0177), which both mean "swelling" in English, also showed a certain level of importance, suggesting that the swelling of mediastinal lymph nodes is involved in the diagnosis of the N classification.

Furthermore, "N2 **キーワード合計**" (0.0232), which is the sum of the frequencies of words for anatomical sites associated with N2 lymph nodes, was confirmed to be a valid indicator in the N2 classification.

On the other hand, "**転移**_frequency2" (0.0224) and "**転移**s_frequency1" (0.0190), meaning both "metastasis" in English, also contributed to the determination of the N classification, indicating that there may be some association between N lesions and distant metastasis (M classification).

Table 3 Feature importances in N classification

	Feature	Importance
1	リンパ節_frequency3	0.047007
2	リンパ節_frequency2	0.037994
3	リンパ_frequency2	0.036443
4	リンパ_frequency3	0.030953
5	腫大_frequency3	0.024422
6	N2キーワード合計	0.023184
7	転移_frequency2	0.022436
8	縦隔_frequency2	0.021020
9	転移_frequency1	0.018997
10	腫大_frequency2	0.017710

4.2.3 M classification

As shown in Table 4, in the M classification prediction, “M1c キーワード合計” (0.0268) showed the highest importance, followed by “腎_frequency3” (0.0259), “転 移_frequency2” (0.0221), “腎_frequency1” (0.0198), “M1c キーワード合計” (0.0190), which mean “kidney,” “metastasis,” “kidney,” and “M1c keyword total” in English, respectively.

The “M1c keyword total” is a feature that sums the frequencies of words for sites suggestive of distant metastasis, such as brain, liver, bone, skin, and adrenal gland, and the results were consistent with the M classification criteria. The frequency of occurrence of “metastasis” (転 移_frequency2, 3) and “multiple” (多 発_frequency1, 2) contributed to the prediction of M classification, reflecting the fact that M1c is a classification involving metastasis to multiple organs.

On the other hand, the relatively high contribution of features related to “kidney” (腎_frequency1, 2, 3) is interesting. This suggests that kidney metastasis is observed at a certain frequency as distant metastasis of lung cancer, but it is less common than liver, bone, or brain metastasis, so further study is needed. In addition, “両側 _frequency2” (0.0123), meaning “bilateral” in English, also showed a certain level of importance, which may reflect the fact that multiorgan metastasis extends to the lungs and other organs on both sides.

Table 4 Feature importances in M classification

	Feature	Importance
1	M1cキーワード合計2	0.026812
2	腎_frequency3	0.025907
3	転移_frequency2	0.022088
4	腎_frequency1	0.019802
5	M1cキーワード合計1	0.018964
6	腎_frequency2	0.018638
7	多発_frequency2	0.017365
8	多発_frequency1	0.016124
9	転移_frequency3	0.015139
10	両側_frequency2	0.012323

4.3 Evaluation results in formal run

The results of the formal run are shown in the bottom row of Table 1. In both fine and coarse, all classification accuracies were markedly lower than the train and validation results. In particular, the accuracy of the T classification was noticeably decreased, suggesting the possibility of overfitting. “Tumor size” is an inherent feature of the T classification model, and this feature could have negatively impacted test accuracy. The accuracy of size estimation in cases where size was not stated and the handling of cases where multiple sizes were stated may have been affected, and more detailed confirmation of the trends in misclassified data is needed. In addition, although “left,” a keyword related to localization, showed a high contribution in the T classification, words related to localization other than “left” (e.g., “right”) are also supposed to appear in the text. Incorporating a variable that aggregates these words, such as “localization,” could improve accuracy.

5 CONCLUSIONS

In this study, we developed highly interpretable models by predefining key terms based on domain knowledge, such as clinical guidelines, and using their frequencies as training data. We obtained results with high medical validity and gained new insights into the contribution of the keyword “kidney” in the M classification model.

This method is characterized by the fact that keywords are used as explanatory variables as they are, resulting in a highly interpretable feature importance table. In addition, this method is versatile and likely equally applicable to any disease for which guidelines exist.

On the other hand, as mentioned above, the problem of overfitting needs to be solved. Furthermore, key terms are expected to be automatedly predefined in the future by analyzing guidelines textually. If these issues are resolved, the method becomes even more useful and valuable.

REFERENCES

[1] A. Fujii *et al.*, "In-hospital countermeasures to prevent forgetting confirmation of diagnostic imaging reports and verification of their effectiveness (in Japanese)," (in Japanese), *Journal of Medical Informatics,* no. 39(Suppl.), pp. 572-574, 2019.

[2] R. Noguchi, K. Torikai, and Y. Saito, "Versatile Data Structuring Framework and Machine Learning Applications for Clinical Utilization of Text Data in Electronic Medical Records," (in Japanese), *Mumps,* vol. 30, pp. 51-57, 2023.

[3] R. Noguchi, "GunNLP at the NTCIR-16 Real-MedNLP Task: Collaborative filtering-based similar case identification method via structured data “case matrix”," *NTCIR 16 Conference: Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies,* pp. 349-352, 2022.

[4] Y. Nakamura *et al.*, "NTCIR-18 RadNLP 2024 Overview: Dataset and Solutions for Automated Lung Cancer Staging," *In Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies,* 2025.

[5] The Japan Lung Cancer Society. "Guidelines for Diagnosis and Treatment of the Lung Cancer (in Japanese)." https://www.haigan.gr.jp/publication/guideline/examination/2024/ (accessed 1.28, 2025).

[6] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Veronika Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," p. arXiv:1706.09516, 2017, doi: 10.48550/arXiv.1706.09516.