Master's Thesis of Data Science

Leveraging Temporal Information for Classification of Lung Cancer's Brain Metastases from Clinical Notes

시간적 맥락을 활용한 임상 기록에서 폐암의 뇌 전이상태 분류

August 2023

Graduate School of DataScience
Seoul National University
Data Science Major

Jiyong An

Leveraging Temporal Information for Classification of Lung Cancer's Brain Metastases from Clinical Notes

Seunggeun Lee

Submitting a master's thesis of
Data Science

July 2023

Graduate School of Data Science
Seoul National University
Data Science Major

Jiyong An

Confirming the master's thesis written by
Jiyong An
July 2023

Chair	Hyung-Sin Kim	(Seal)
Vice Chair	Seunggeun Lee	(Seal)
Examiner	Jay-Yoon Lee	(Seal)

Abstract

Lung cancer is one of the most common types of cancer that frequently metastasizes to the brain. For optimal patient care and informed decision-making, accurate metastatic status classification is crucial.

In this study, we propose two approaches that can leverage temporal information with contextual information in clinical notes to categorize cancer status into four distinct classes. First, we combined a BERT-based model with a Conditional Random Field (CRF) layer. Second, we built a Bidirectional Long Short-Term Memory (BiLSTM) model with sequences of word embedding from the pre-trained model. The dataset comprises 13,684 clinical notes, of which only 606 are annotated.

We first fine-tune ClincalBERT with 450 annotated data, achieving an accuracy of 73.9 %. To augment the model's performance, a CRF layer is integrated on top of fine-tuned ClincalBERT, exploiting the temporal information provided by each note's date. The CRF layer is trained using 4,237 pseudo-labeled notes with a confidence threshold of 0.95, resulting in a model with 89.1 % accuracy. Additionally, we employ a semi-supervised approach while training a BiLSTM model with Clinical BERT's word embeddings, resulting in a model with 93.4 % accuracy.

Our findings underscore the significance of leveraging longitudinal information and semi-supervised learning techniques for cancer status classification using clinical notes, with implications for personalized medicine and clinical support systems.

Keyword: clinical note, pseudo labeling, semi-supervised learning, conditional random field, BERT, LSTM

Student Number: 2021-27031

Table of Contents

Chapter 1. Introduction

1.1. Study Background

Lung cancer is one of the leading causes of cancer-related deaths worldwide. It is notorious for its high rate of metastasis, or spread, to other organs, particularly the brain. In fact, lung cancers account for over 60% of all brain metastases [1]. Metastatic spread to the brain not only significantly deteriorates patients' quality of life but also reduces overall survival rates, emphasizing the need for accurate detection for informed treatment planning.

Current detection and diagnosis methods, such as microscopic examination or conventional imaging techniques, such as magnetic resonance imaging (MRI) and computed tomography (CT) scans, although crucial, are not always conclusive [2]. Moreover, these approaches often encounter issues of subjectivity and interpretation inconsistencies, potentially compromising the accuracy and effectiveness of cancer diagnoses [3].

In recent years, there has been a growing interest in using computational techniques to automatically classify cancer metastasis from clinical notes [4, 5, 6]. Clinical notes are a rich source of data that can be used to identify cancer metastasis, but manual extraction and analysis of this data are laborious and prone to human error. Therefore, the utilization of advanced computational techniques for automatic metastatic status classification is not only beneficial but necessary.

1.2. Purpose of Research

At Seoul National University Bundang Hospital (SNUBH), there is an increasing need to ascertain the status of brain metastasis in lung cancer patients for accurate cancer staging. This information is vital for proper recording of the patient's cancer staging status within their database. To address this, our study aims to assist the SNUBH medical team by providing a reliable computational model for metastasis classification from clinical reports.

The advent of Electronic Health Record (EHR) systems and advancements in natural language processing (NLP) have led to a greater utilization of clinical notes. These notes encapsulate rich information pertaining to patients' medical histories, diagnoses, and treatment responses [7]. This shift towards data-driven decision-making significantly aids in improving patient care.

One of the most promising NLP models is Bidirectional Encoder Representations from Transformers (BERT). BERT has been shown to perform well on a variety of NLP tasks, including text classification, question answering, and natural language inference [8]. More specifically, in the healthcare sector, ClinicalBERT, a domain-specific adaptation of BERT, has exhibited remarkable capability in processing and understanding medical language due to its pre-training on a large corpus of clinical text [9].

Despite the promising results exhibited by pre-trained models, their ability to capture the temporal structure within EHRs remains limited. This is a significant limitation, as temporal information can be crucial for accurate classification of brain metastases status as there are strong dependencies for the patients' brain metastases status over time. Therefore, the present study seeks to address this gap by

incorporating temporal information into clinical notes. This will provide additional insights that could aid in the accurate classification of brain metastases status.

1.3. Objectives

As depicted in Figure 1, the principal aim of this thesis is to develop an accurate classification system capable of discerning brain metastases statuses from the clinical notes of lung cancer patients. This system will integrate both contextual and temporal information, leveraging the unique strengths of these two data sources. By doing so, this research has the potential to contribute to enhance the accuracy of cancer diagnosis and prognosis. Ultimately, it is anticipated that these advancements will contribute to improvements in patient care and outcomes.

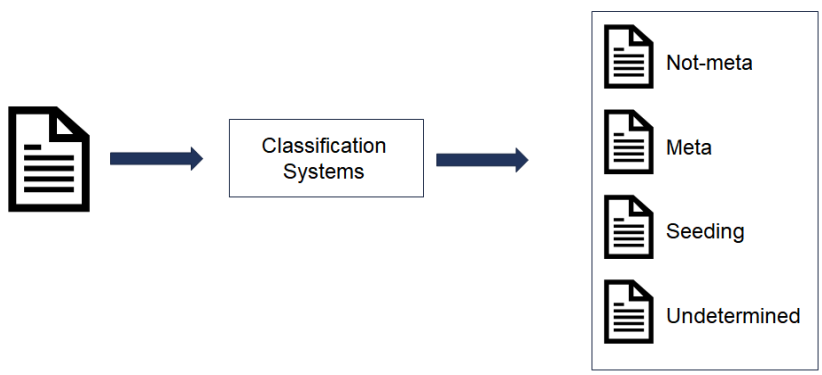

Figure 1. Objective of the study

Chapter 2. Preliminary Literature Review

2.1. Deep learning for text classification

Research in deep learning for text classification can be broadly categorized into two distinct areas. The first concentrates on the development of word embedding models to represent words as vectors in a high-dimensional space. This allows words with similar meanings to be represented by similar vectors, and the quality of these embeddings has a significant impact on the outcome of text classification tasks. Two popular word embedding models are Word2Vec [10] and BERT. Word2Vec uses either the Skip-gram or Continuous Bag of Words (CBOW) algorithm to learn word embeddings. BERT uses a self-attention mechanism to learn word embeddings that are more contextually aware.

The second area pertains to deep neural networks, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory Networks (LSTMs). Deep neural networks are a type of machine learning model that can learn complex patterns from data. They are often used for text classification tasks because they can learn to represent the semantics of text documents. For instance, Kim employed a CNN for sentence classification [11], representing sentences or documents as matrices with each row corresponding to the vector representation of words or characters. The text representation was then learned through convolution and pooling operations. Luo utilized an RNN to learn text representations by splitting sentences into sequences of words and letting the RNN model learn from these sequences [12].

One of the notable experiments in this field was conducted by Senders et al., wherein they undertook a comparative analysis of different NLP models aimed

at the automated classification of medical text [13]. Their data source was magnetic resonance imaging (MRI) brain reports from patients with brain metastases. The reports were manually annotated to provide a binary classification: whether a patient had a single metastasis or two or more metastases. Among the six diverse NLP models that they evaluated, the Long Short-Term Memory (LSTM) model demonstrated the most promising results. When trained on a data set comprising 1,179 reports, this LSTM model achieved an accuracy rate of 87% on the binary classification tasks, thereby underscoring the potential of NLP in the classification of medical text.

2.2. Improving BERT's performance with CRF layer

In the pursuit of improving the BERT model, various methods have been proposed by researchers. Notably, Souza et al. introduced an additional layer to the BERT model for better sequence classification [14].

Even though standard BERT-based classifiers have shown promising results in various NLP tasks, they face a significant challenge when applied to sequential decision-making [14]. While they process each sequence point using information from both preceding and succeeding points, the final classification decision for a given point is made independently of the decisions at the other points. This means that while they consider the context of neighboring data points for understanding each point, they do not utilize the classification decisions of these neighboring points. This can lead to less accurate results, especially in scenarios where the context of classification decisions in the sequence is important.

Addressing this issue, Souza et al. devised a model that combines BERT

with a token-level classifier and further supplements it with a Linear-Chain Conditional Random Field (CRF) [15]. The CRF serves as a labeler, with the label assigned to a current word being influenced by the label of the previous word. By integrating this additional CRF layer, the classifier can consider the classification decisions from adjacent points in the sequence before reaching its own decision. This improvement allows for more accurate and contextually aware sequence classification outcomes.

2.3. BiLSTM

The BiLSTM model constitutes two separate LSTM structures, processing the input sequences in both forward and reverse orders, respectively. This bidirectional approach effectively encapsulates the features of the input sequence. The final word vector representation is obtained by concatenating the output vectors from the two LSTM networks, serving as the ultimate feature descriptor for the sentence.

As illustrated in Figure 2, an LSTM unit comprises three gates: the forget gate, input gate, and output gate [17]. The forget gate governs the degree of information to retain from the previous cell state, deciding which information to discard. The input gate modulates the extent of new input information added to the cell state, while the output gate determines the update of the current memory state and the output of the hidden layer.

The LSTM employs the forget gate to determine what information the cell state should preserve. The input gate determines how much new input information should be saved to the cell state, and the output gate decides the final output information. The following equations define the operations of an LSTM unit at the t-th timestep:

$$Forget\ Gate(f_t) = \sigma\big(W_f\,[h_{t-1}, x_t] + b_f\big),$$

$$Input\ Gate(i_t) = \sigma(W_i\,[h_{t-1}, x_t] + b_i),$$

$$Cell\ State(c'_t) = \tanh(W_c\,[h_{t-1}, x_t] + b_c),$$

$$\text{Candidate } Cell\ State(c_t) = f_t * c_{t-1} + i_t * c'_t,$$

$$Output\ Gate(o_t) = \sigma(W_o\,[h_{t-1}, x_t] + b_o),$$

$$Hidden\ State(h_t) = o_t * tanh(c_t)$$

where f_t , i_t , and o_t represent the forget gate, input gate, and output gate, respectively, h_{t-1} , W, and b are the output of the previous hidden layer state, weight, and bias of gate neurons, and c'_t and c_t are the cell state and the candidate of cell state.

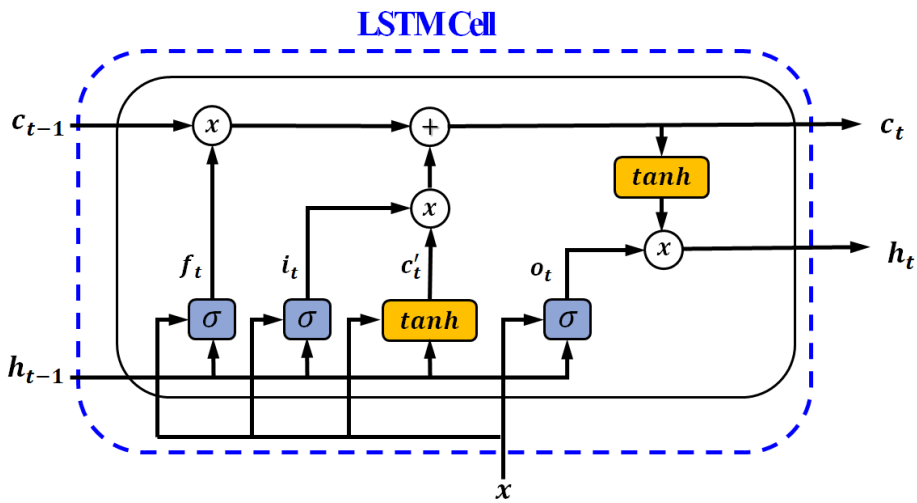

Figure 2. LSTM Architecture

While the LSTM model addresses the issue of short-term memory by regulating information flow through internal gate mechanisms, it only processes information from front to back, which may limit comprehensive semantic capture. The BiLSTM model, however, comprises two LSTMs that effectively encapsulate

bidirectional semantics. One LSTM processes the sequence in the forward direction while the other does so in reverse, and the outputs of the two LSTMs are combined. The hidden state $h_t \ at$ t-th position in the BiLSTM model, where $\overrightarrow{h_t} \ and \ \overleftarrow{h_t}$ are the forward hidden layer state and the backward hidden layer state, respectively.

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t}$$

Chapter 3. Methodology

We utilize two separate methods to classify brain cancer metastasis from lung cancer patients' clinical reports: the Conditional Random Fields (CRF) layer added to the fine-tuned ClinicalBERT, and a bidirectional Long Short-Term Memory (Bi-LSTM) model trained with ClinicalBERT's sentence-level embedding, which is extracted from [CLS] token. Each method takes advantage of a pre-trained model's capacity for contextual understanding and is designed to handle sequential information.

3.1. Data Preparation

Our dataset comprises 13,684 clinical notes, of which only 606 have been annotated. These selected reports were manually reviewed and annotated by domain experts from the Seoul National University Bundang Hospital (SNUBH). From the annotated subset, 500 notes were acquired through random sampling, while the remaining 106 notes were specifically selected from patients who have generated multiple records over time. This method of selection ensures that our dataset encompasses the sequential progression of cancer status within individual patients. Consequently, this provides a rich contextual background for the development and evaluation of our proposed cancer status classification models, potentially leading to more accurate and insightful results.

As demonstrated in Figure 3, the notes were grouped by PERSON_ID and ordered by NOTE_DATE to model the temporal feature. We define each patient's record cluster as a "sequence", and in this study, we only consider sequences of length ten or more. Table 1 provides a description of the number of sequences in annotated notes and unlabeled notes, and Table 2 provides the distribution of labels

in the annotated dataset.

NOTE_ID	PERSON_ID	NOTE_TEXT	NOTE_DATE	LABEL
0	100	Two tiny enhancing nodular lesions at right occipital and left parietal cortex - > probable metastases	2016-03-30	Meta
1	100	Reappeared another tiny metastasis in left parietal cortex, different site from 2016-03-30 MRI	2017-02-27	Meta
2	100	No change or sl. less decreased size of a small ring enhancing metastasis in right occipital lobe	2017-06-18	Meta
3	100	Further decreased size of multiple enhancing metastases in both cerebral hemispheres and both cerebellum. [Finding] NC of tiny enhancing lesion at left IAC. NC of mild brain atrophy	2018-01-02	Meta
4	101	Compatible with multiple brain metastasis. 환자이름: 홍길동 Lung cancer 환자임. 1cm 미만의 크기를 갖는 4개의 small nodular enhancing lesion.....	2018-02-27	Meta
5	101	Disappeared four small nodular enhancing lesion at both cerebellum and left occipital cortex. No new metastasis.	2018-03-13	Not-meta
6	102	Focal linear and tiny nodular enhancements at right precentral gyrus surface -> r/o metastasis, possible seeding	2017-03-23	Undetermined
7	102	Still no evidence of new enhancing metastasis in the brain. No change in the following findings : -about 2cm sized enhancing	2017-05-30	Not-meta
8	102	Still no new metastasis in the brain. Slowly growing 2cm sized enhancing lesion is right slide pituitary gland, invading right cavernous	2017-09-04	Not-meta
...	...	...	...	...

Figure 3. Example of Dataset with sequences

	Number of Notes	Number of notes with sequences	Number of sequences
Un-labeled	13,078	4,511	276
Randomly sampled	500	0	0
Selectively sampled	106	106	10
Total	13,684	4617	286

Table 1. Data Description

Labels	Randomly sampled	Selectively sampled
Not-meta	242	45
Meta	105	49
Seeding	73	1
Undetermined	80	11
Total	**500**	**106**

Table 2. Label distribution of annotated dataset

3.2. Fine-tuning ClinicalBERT

In the first stage of our methodology, we fine-tuned the base model, ClinicalBERT,

using a subset of our annotated data. We selected 450 notes from the 500 randomly sampled notes for this purpose. The remaining 50 randomly sampled annotated notes were reserved for validation purposes.

Fine-tuning is a common practice in transfer learning, wherein a pre-trained model is further trained on a specific task using a smaller dataset. This process allows the model to adapt its generalized knowledge learned from the large pre-training corpus to the specific task at hand. This fine-tuning procedure was carried out to ensure that ClinicalBERT is more suited to our specific task and can leverage the intricacies and nuances present in our dataset.

3.3. ClinicalBERT with CRF layer

3.3.1 Data preparation with pseudo-labeling

Figure 4 shows the overall research procedure for combining CRF layer with fine-tuned ClinicalBERT. Given the limited number of annotated sequences available, we augmented the training data by using pseudo-labeled notes generated by the fine-tuned ClinicalBERT. In an effort to enhance robust training, we selected notes that possessed an estimated probability exceeding 0.95. After this augmentation process, the CRF layer was trained with 248 sequences of notes, amounting to a total of 4,237 notes.

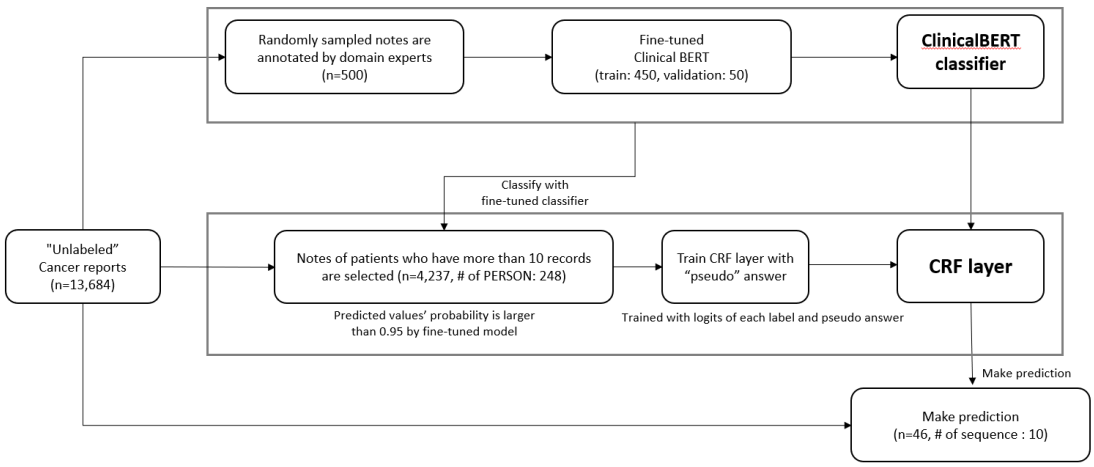

Figure 4. ClinicalBERT with CRF layer workflow

3.3.2 Condtional Random Field (CRF) layer

While Fine-tuned ClinicalBERT has achieved significant success in relatively simple tasks such as text classification, it is constrained when faced with tasks where output labels have strong interdependencies. One such task is brain metastasis classification, where metastases status exhibits strong dependencies over time. Hence, instead of independently modeling classification decisions, we jointly model them using a Conditional Random Field (CRF).

For an input sentence $X = (x_1, x_2, \dots, x_n)$, we define P as the emission scores generated by the ClinicalBERT classifier. P is of size $n \ x \ k$, where n is the number of notes in the sequence and k is the number of distinct classes. $P_{i,j}$ denotes the score of the j^{th} label of the i^{th} note in a sequence. For a sequence of predictions $y = (y_1, y, \dots, y_n)$, its score, denotes as s, is defined by the following equation:

$$s = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_{i+1}}$$

Here, A is a transition score matrix where $A_{i,j}$ represents the score of transition from the label i to label j. Considering that there are four classes in our study the size of

matrix is 4 x 4.

Assuming a total of N possible paths, the softmax over all possible sequences of metastatic status yields a probability for the sequence y:

$$p(y|X) = \frac{e^{s(X,y)}}{\sum_1^N e^{s(X,y)}}$$

During training, tge log-probability of the correct label sequence is maximized, as expressed in the equation:

$$LogLossFunction = log\frac{e^{s(X,y)}}{\sum_1^N e^{s(X,y)}}$$

By adding a negative sign, we transform our model to learn by miminizing the loss function.

$$LogLossFunction$$

$$= -log\frac{e^{S_{RealPath}}}{e^{S_1} + e^{S_2} + \ ... \ + e^{S_N}}$$

$$= -(log(e^{S_{RealPath}}) - (log(e^{S_1} + e^{S_2} + \ \dots \ + e^{S_N}))$$

$$= -(S_{RealPath} - log(e^{S_1} + e^{S_2} + \ \dots \ + e^{S_N}))$$

Through this formulation, the CRF produce a valid sequence of output labels. Finally, CRF layer leverage Viterbi algorithm [16] the CRF layer obtain the sequence with the maximum score.

Figure 5 provides a visual representation how this method works. (1) First, the fine-tuned ClinicalBERT takes a sequence of notes as input and (2) generates emission scores, which are logits for each class. (3) These emission scores and corresponding labels form the the inputs of the CRF layer, which calculates transition scores for each transition among the label. Figure 6 shows the transition score matrix

that our CRF layer computed. (4) Using the emission score and transition scores, the CRF layer computes "path scores" for all possible label sequences given an input sequence. (5) The sequence with the highest path score is identified using viterbi algorithm.

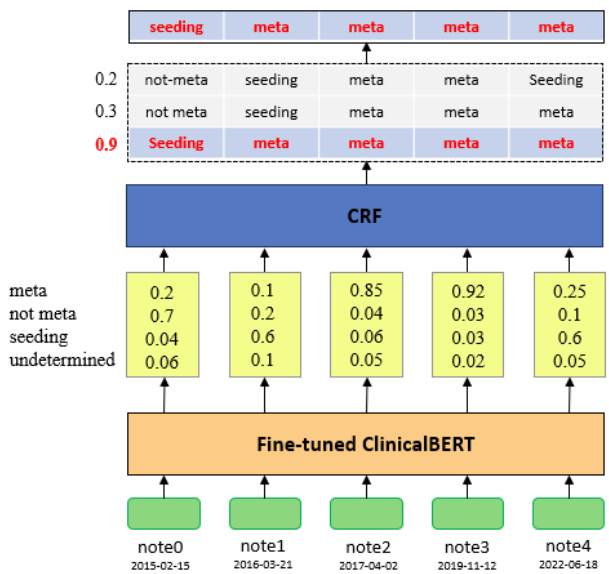

Figure 5. Process of ClinicalBERT with CRF layer

		TO			
		no meta	meta	seeding	undetermined
FROM	no meta	0.66	0.13	-0.12	0.19
	meta	-0.13	0.87	0.23	-0.32
	seeding	-0.63	0.21	0.94	-0.20
	Undetermined	-0.58	-0.17	-0.39	0.21

Figure 6. Transition Score Matrix

3.4. BiLSTM approach

3.4.1 Semi-supervised learning with self-training method

While the ClinicalBERT classifier with a CRF layer effectively models label

dependencies, it does not fully capture contextual dependencies in the input data. This is particularly crucial as cancer progression can vary significantly among individuals, making it essential to consider the contextual dependencies within a sequence of clinical notes.

To address this challenge, we employ a BiLSTM model trained with ClinicalBERT's sentence level word embeddings and their corresponding labels. The overall workflow for the BiLSTM approach is depicted in Figure 7.

Initially, we began by partitioning 106 selectively annotated records into a training set and test set. We selected the first six records from each individual as the training set, and the remaining 46 notes were allocated to the test set. However, given the limited number of annotated notes available for training, we employed a semi-supervised learning technique known as self-training approach. The process of self-training is illustrated in Figure 8. The self-training approach involves two key steps:

1. The model is initially trained solely on the annotated data.
2. With this trained model, we generate pseudo-labeled data for the unlabeled instances. This is achieved by assigning labels to the unlabeled instances with a confidence threshold of 0.95.

This semi-supervised learning approach helps us to effectively use unlabeled data by leveraging the model's ability to infer labels for instances not in the annotated dataset. However, the self-training method presents a potential pitfall known as the 'overconfidence problem', where the model becomes too certain about its predictions on the unlabeled data, leading to a bias towards its initial predictions. This is counterproductive as the initial predictions could be erroneous and the model ends up reinforcing these mistakes.

To mitigate this overconfidence problem, we adjust the confidence

threshold after the first 10 epochs, reducing it from 0.95 to 0.90. By lowering the threshold, we introduce a degree of uncertainty into the model's predictions on unlabeled data. This uncertainty can help the model explore different possibilities and learn better from the newly pseudo-labeled data in the subsequent epochs of training."

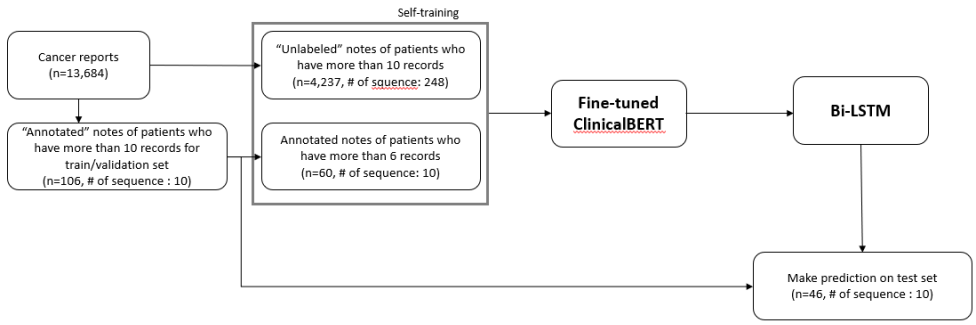

Figure 7. Workflow of BiLSTM approach

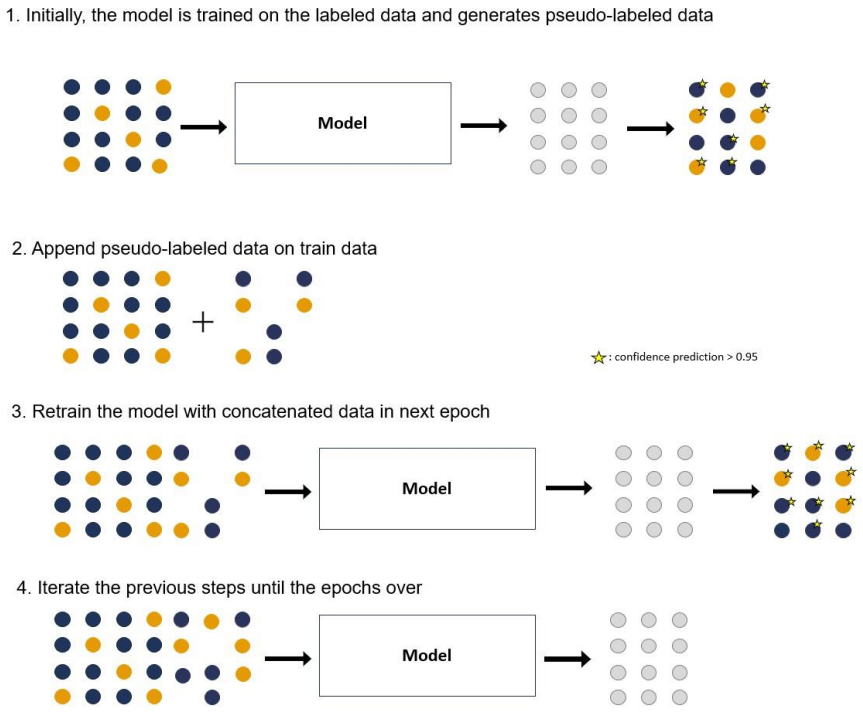

Figure 8. Self-training method

3.4.2 Train BiLSTM with ClinicalBERT's sentence level embedding

In this study, we enhance the understanding of a patient's condition and the relevance of specific information within each note by integrating both past and future contexts using a Bidirectional Long Short-Term Memory (BiLSTM) model.

Our process, as illustrated in Figure 9, initiates with ClinicalBERT, which has been fine-tuned to receive a sequence of clinical notes as input and generate word embeddings for each note. Given that the [CLS] token encapsulates the overall information of a sentence, we extract its word embedding to represent each note on a broader level. These sentence-level word embeddings are then fed into the BiLSTM model. By integrating the fine-tuned ClinicalBERT's embeddings, we expect to bolster the model's robustness and its ability to generalize, which could enhance its performance even when the available data is limited.

By taking into account the contextual information and labels in a bidirectional manner, BiLSTM is able to decipher complex patterns and provide accurate predictions that account for both temporal and contextual dependencies. This strategy elevates the model's capacity to identify unique features associated with each individual's cancer progression, thereby potentially increasing the model's accuracy on brain metastases status classification."

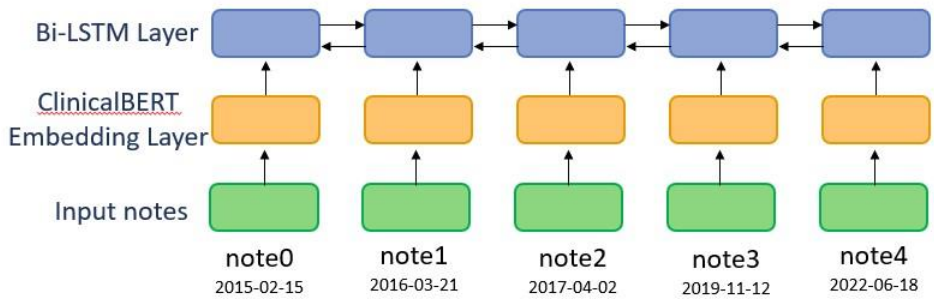

Figure 9. Process of BiLSTM approach

Chapter 4. Results and Discussion

4.1. Results

The performance of each model was evaluated using a set of 46 selectively annotated notes. These notes included 21 instances of no metastases, 24 instances of metastases, and one instance of an undetermined status. Notably, no instances of 'seeding' were included in these test notes.

Table 3 presents a comparative analysis of precision and recall across the three models. The results indicate that the proposed methods generally outperform the base model in terms of both precision and recall. An exception to this is observed in the case of 'not meta', for which the base model achieved the highest precision. However, this appears to be primarily due to its relatively low recall.

The Table 4 shows the comparison of accuracy for the three models. Compared to the base model, both CRF and BiLSTM methods achieved considerably improved accuracy, increased by 15.2 percent and 19.5 percent, respectively.

Test data	**Base**		**CRF**		**Bilstm**	
	Precision	`Recall	Precision	Recall	Precision	Recall
Not meta	**1.00**	0.83	0.83	0.95	0.89	**1.00**
Meta	0.83	0.71	**1.00**	**0.88**	**1.00**	0.86
Undetermined	0	0	0	0	**1.00**	**1.00**

Table 3. Comparison of precision and recall

	Correct	**Incorrect**	**Accuracy**
ClinicalBERT	34	12	73.9
ClinicalBERT + CRF layer	41	5	89.1
ClinicalBER + BiLSTM	43	3	93.4

Table 4. Comparision of accuracy

4.2. Discussion

In this study, we presented two distinct approaches that effectively employ both temporal and contextual information from clinical notes to classify cancer status into four separate categories. A considerable improvement in performance was observed between our base model and the enhanced models—ClinicalBERT with a CRF layer, and the BiLSTM model. This result shows the significance of leveraging temporal information for metastatic status classification.

However, we acknowledge that this disparity could be attributed to the difference in data distribution between the training and test sets, as demonstrated in Table 2. The test set, comprising selectively sampled notes, predominantly includes either metastatic or non-metastatic cases. However, the randomly sampled notes in the training set encompass a higher proportion of 'seeding' and 'undetermined' cases. Upon validating our base model with 50 randomly sampled notes, we observed a significant increase in accuracy to 86%.

We acknowledge that our results warrant further validation using a more comprehensive set of test notes. It is also crucial to evaluate the robustness and generalizability of our models by applying them to clinical notes from various hospitals.

As a further avenue of research, it would be insightful to investigate the specific sections or clauses within the notes that our model prioritizes when making its decisions. Enhancing the model's explainability could provide deeper insights into its decision-making process. Importantly, a higher degree of explainability can foster trust among medical experts by making the model's decision-making process more transparent and comprehensible. This comprehension could consequently

highlight potential areas for model refinement, thereby leading to even more accurate predictions.

Appendix

Figure 10 illustrates the manner in which the CRF layer refines ClinicalBERT's predictions by taking into account temporal dependencies. Notably, the model is capable of distinguishing the "undetermined" metastasis status into specific categories, demonstrating the value of incorporating temporal information.

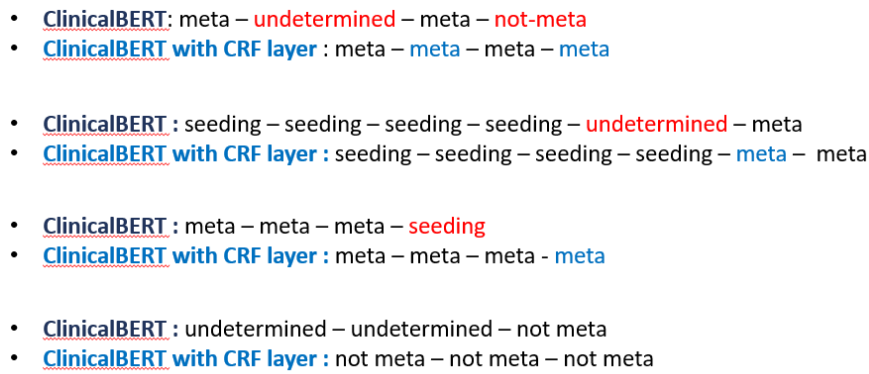

Figure 10. Sample results from CRF approach

Bibliography

1. K. M. J. K. H.-J. I. J.-S. S. H. G. Jung Kyu-Won, Won Young Joo. Prediction of cancer incidence and mortality in korea, 2022. crt, 54(2):345–351, 2022. doi: 10.4143/crt.2022.179.
2. Fink, Kathleen, and James Fink. "Imaging of brain metastases." *Surgical neurology international* 4 (2013): 209.
3. K.-Y. Su and W.-L. Lee. Fourier transform infrared spectroscopy as a cancer screening and diagnostic tool: A review and prospects. Cancers, 12(1), 2020.
4. Do, Richard KG, Kaelan Lupton, Pamela I. Causa Andrieu, Anisha Luthra, Michio Taya, Karen Batch, Huy Nguyen et al. "Patterns of metastatic disease in patients with cancer derived from natural language processing of structured CT radiology reports over a 10-year period." *Radiology* 301, no. 1 (2021): 115-122.
5. Yang, Ruixin, Di Zhu, Lauren E. Howard, Amanda De Hoedt, Stephen B. Williams, Stephen J. Freedland, and Zachary Klaassen. "Identification of Patients With Metastatic Prostate Cancer With Natural Language Processing and Machine Learning." *JCO clinical cancer informatics* 6 (2022): e2100071.
6. Banerjee, Imon, Selen Bozkurt, Jennifer Lee Caswell-Jin, Allison W. Kurian, and Daniel L. Rubin. "Natural language processing approaches to detect the timeline of metastatic recurrence of breast cancer." *JCO clinical cancer informatics* 3 (2019): 1-12.
7. Abul-Husn, Noura S., and Eimear E. Kenny. "Personalized medicine and

the power of electronic health records." *Cell* 177, no. 1 (2019): 58-69.

8. J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.
9. E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott. Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323, 2019.
10. Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).
11. Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.
12. Luo Y. 2017. Recurrent neural networks for classifying relations in clinical notes. Journal of Biomedical Informatics, 72, 85-95.
13. Senders JT, Karhade AV, Cote DJ, Mehrtash A, Lamba N, DiRisio A, Muskens IS, Gormley WB, Smith TR, Broekman MLD, Arnaout O. Natural Language Processing for Automated Quantification of Brain Metastases Reported in Free-Text Radiology Reports. JCO Clin Cancer Inform. 2019 Apr;3:1-9. doi: 10.1200/CCI.18.00138. PMID: 31002562; PMCID: PMC6873936.
14. Souza, Fábio, Rodrigo Nogueira, and Roberto Lotufo. "Portuguese named entity recognition using BERT-CRF." arXiv preprint arXiv:1909.10649 (2019).
15. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields:

Probabilistic models for segmenting and labeling sequence data. 2001.

16. G. D. Forney, "The viterbi algorithm," in *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268-278, March 1973, doi: 10.1109/PROC.1973.9030.
17. Pisa, Ivan & Santín, Ignacio & Morell, Antoni & Lopez Vicario, Jose & Vilanova, Ramon. (2019). LSTM-Based Wastewater Treatment Plants Operation Strategies for Effluent Quality Improvement. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2950852.
18. Yu, Yong, et al. "A review of recurrent neural networks: LSTM cells and network architectures." *Neural computation* 31.7 (2019): 1235-1270.

Abstract

폐암은 뇌로 자주 전이되는 가장 흔한 암 유형 중 하나로, 최적의 환자 치료와 정보 기반 의사 결정을 위해서는 암세포의 뇌 전이 상태를 정확히 분류하는 것이 중요하다. 본 연구에서는 폐암 환자의 MRI 판독소견서의 시간적 정보와 맥락적 정보를 함께 활용하여 폐암의 뇌전이 상태를 분류할 수 있는 두 가지 접근법을 제안한다. 첫번째 방법으로는 BERT기반의 사전된 모델을 fine-tuning하여 Conditional Random Field (CRF) 레이어와 결합하였으며, 두번째 방법으로는 사전학습된 모델에서 문장 수준의 임베딩 시퀀스를 추출하여 Bidirectional Long-Short Term Memory (BiLSTM)모델을 구축하였다. 데이터셋은 총 13,684개의 임상기록으로 구성되어있으며, 이 중 606개의 데이터만이 주석처리 되었다. 주석처리된 데이터의 수가 부족한 문제는 준지도학습 방법론을 동원하여 해결을 시도하였다. 450개의 주석이 달린 데이터를 활용하여 ClinicalBERT를 fine-tuning하였으며, 이를 통해 73.9%의 정확도를 달성하였다. 모델의 성능을 향상시키기 위해, 미세조정된 ClinicalBERT위에 CRF 레이어를 통합하였고, 이는 89.1%의 정확도를 달성하였다. 마지막으로, ClinicalBERT의 문장 수준의 임베딩을 사용하여 BiLSTM을 학습시켜 93.4%의 정확도를 달성하였다.

우리의 연구 결과는 임상기록을 사용한 폐암의 뇌 전이 상태 분류를 위해 시간 정보 준지도 학습 기법을 활용하는 것의 중요성을 확인하였고, 보다 신뢰할 수 있는 모델을 제공함으로써 의료진의 의사 결정을 도울수 있음을 시사한다.

주요어: 의료노트, 슈도-레이블링, 준지도학습, 조건부 랜덤 필드, BERT, LSTM

Student Number: 2021-27031