J Korean Med Sci. 2026 Mar 2;41(8):e79
https://doi.org/10.3346/jkms.2026.41.e79
eISSN 1598-6357·pISSN 1011-8934

JKMS

Original Article
Medical Informatics

Check for updates

Application of a Natural Language Processing Framework for Data Extraction From Pathology Reports Across Multiple Cancer Types

Phillip Park,[1,2] **Yeonho Choi**,[2] **Nayoung Han**,[3*] **Soobin Park**,[2] **Ye-Lin Park**,[2] **Juyeon Hwang**,[2,4] **Kui Son Choi**,[2,5] **Chong Woo Yoo**,[3] **and Hyun-Jin Kim** [2]

[1]Department of Digital Health, Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University, Seoul, Korea
[2]Cancer Data Center, National Cancer Control Institute, National Cancer Center, Goyang, Korea
[3]Department of Pathology, National Cancer Center, Goyang, Korea
[4]Department of Health Informatics and Biostatistics, Graduate School of Public Health, Yonsei University, Seoul, Korea
[5]Graduate School of Cancer Science and Policy, National Cancer Center, Goyang, Korea

Received: Feb 18, 2025
Accepted: Jun 22, 2025
Published online: Feb 9, 2026

Address for Correspondence:
Hyun-Jin Kim, PhD
Cancer Data Center, National Cancer Control Institute, National Cancer Center, 323 Ilsan-ro, Ilsandong-gu, Goyang 10408, Republic of Korea.
Email: hyunjin@ncc.re.kr

*Current affiliation: Department of Pathology and Translational Genomics, Samsung Medical Center, Seoul 06351, Korea.

ORCID iDs
Phillip Park
https://orcid.org/0000-0002-8769-6383
Yeonho Choi
https://orcid.org/0009-0007-4296-6451
Nayoung Han
https://orcid.org/0000-0001-8710-4280

ABSTRACT

Background: Pathological reports provide comprehensive insights into the clinical and pathological features of different cancer types. However, extraction of this semi-structured data for research is challenging. To better utilize pathology reports in cancer studies, we developed an efficient natural language processing (NLP) system to automate the extraction of items from pathology reports, facilitating streamlined storage, retrieval, and analysis of clinical data in a centralized database.

Methods: To determine the optimal model for our study, we conducted a comparative analysis of various deep learning architectures, including long short-term memory, convolutional neural network, and transformer-based models such as bidirectional encoder representations from transformers (BERT), BioBERT, and ClinicalBERT. The proficiency of the ClinicalBERT model in medical terminology and context significantly enhanced the accuracy and efficiency of data extraction from these reports.

Results: Among the aforementioned models, ClinicalBERT exhibited the best performance and was selected as the base model. The ClinicalBERT model demonstrated an exceptional performance in accurately classifying variables across multiple cancer types. Regarding stomach cancer, F1 scores (F1 = 1.0) were achieved for variables such as angiolymphatic invasion, and operation name (F1 = 1.0); however, a lower performance was observed for distant metastasis (F1 = 0.3889). Regarding liver cancer, high performance was consistently observed for most variables, with F1 scores above 0.99. Regarding colorectal cancer, F1 scores were achieved for variables such as Dworak's grade, lymph node, operation name, and total mesorectal excision (F1 = 1.0), while slightly lower but acceptable performance was noted for surgical margin (F1 = 0.9259). Regarding breast cancer, F1 scores were achieved for several variables including nipple margin, organ, and superficial margin (F1 = 1.0), while strong performances were noted for lateral and medial margins (F1 > 0.94).

Conclusion: This study underscores the efficacy of NLP systems, specifically the ClinicalBERT model, in automating the extraction of important clinical data from pathology

Soobin Park
https://orcid.org/0009-0008-0955-8998
Ye-Lin Park
https://orcid.org/0009-0007-5552-4847
Juyeon Hwang
https://orcid.org/0009-0004-5571-2168
Kui Son Choi
https://orcid.org/0000-0001-5336-3874
Chong Woo Yoo
https://orcid.org/0000-0002-5221-4516
Hyun-Jin Kim
https://orcid.org/0000-0003-4160-4815

Funding

This study was supported by a grant from the National Cancer Center (grant No. 2310400). The funders had no role in study design, data collection and analysis, decision.

Disclosure

The authors have no potential conflicts of interest to disclose.

Data Sharing Statement

The datasets generated and/or analyzed during the current study contain potentially sensitive information or risks of personally identifiable information. Data access requests can be submitted through the Data Management Committee of National Cancer Center via email (data@ncc.re.kr). Requests will be evaluated based on the researcher's study design, intended usage environment, and the level of data pseudonymization, ensuring both scientific value and data protection compliance.

Author Contributions

Conceptualization: Park P, Kim HJ. Data curation: Park S, Han N, Yoo CW. Formal analysis: Choi Y. Funding acquisition: Kim HJ. Investigation: Choi Y. Methodology: Choi Y, Park YL. Project administration: Kim HJ. Resources: Kim HJ. Software: Choi Y, Park YL. Supervision: Choi KS, Kim HJ. Validation: Han N, Yoo CW. Visualization: Choi Y, Park YL. Writing - original draft: Park P. Writing - review & editing: Park P.

reports across various cancer types. This approach can not only simplify the process but also enhance the accuracy of the extracted information.

Keywords: Pathology Report; Natural Language Processing; Cancer; Database

INTRODUCTION

Pathology reports contain a wealth of information regarding biological and clinical features, histological types, and the tumor, node, metastasis staging systems of cancers, which are necessary to make optimal treatment decisions and provide valuable insights into the clinical and pathological features of patients. However, utilizing this information can be challenging due to its semi-structured nature as research on clinical oncology and cancer requires structured data.[1] In Korea, the use of unstructured data for research has traditionally been challenging due to privacy concerns. However, in 2024, the Health and Medical Data Utilization Guidelines were updated to allow the use of unstructured data structured through natural language processing (NLP) techniques, marking a significant advancement in medical data utilization.[2]

Traditionally, pathology reports are parsed using regular expressions, which are labor-intensive and require clinical experts or pathologists.[3] As the complexity between patterns increases with the number of rules, regular expressions become ineffective when the documents structures change.[4-6]

In recent years, machine and deep learning approaches have been increasingly used to extract information from pathology reports. The most widely used machine learning methods for this purpose include support vector machines, followed by naïve Bayes, conditional random fields, and random forests.[7,8] Deep learning approaches, such as long short-term memory (LSTM) and convolutional neural network (CNN), are becoming increasingly popular in medical research, particularly for named entity recognition in biomedical contexts.[9-11] The bidirectional encoder representations from transformers (BERT) model is another highly powerful deep learning NLP algorithm.[12] Various NLP methods have been applied to extract information from pathology reports, including models like BioBERT, a domain-specific language representation model pre-trained on a large-scale biomedical corpus.[13] Additionally, ClinicalBERT, which is specifically trained on clinical notes and medical records, has shown promising results in understanding medical terminology and context within pathology reports.[14]

Although previous studies have structured pathology reports using NLP, the number of extractable variables remains limited.[15-17] Previous studies have extracted three to four variables, such as pathology grade, lesion location, and treatment method from pathology records, extracted biomarker information through the lymphoma classification tool, or extracted immunopathology information for a specific purpose.[15]

The centralization and structuring of pathology reports represent critical advancements in medical informatics and clinical research. Such systematic organization enhances data quality and consistency through standardized documentation protocols, enabling more robust analyses; facilitates seamless data sharing and collaboration across medical institutions and research facilities, promoting multicenter studies and collaborative research

initiatives; and provides researchers with streamlined access to clinical information, significantly reducing the time and resources required for data collection and analysis.

Therefore, this study aimed to develop an efficient and precise NLP system to completely automate and streamline the process of extracting vital clinical data from pathology reports. By automating this process, the extracted data can be seamlessly integrated into a centralized database to ensure efficient and dependable storage, retrieval, and analyses. Finally, a framework was designed to provide researchers with essential clinical information.

METHODS

Overview

Two distinct approaches exist in the field of pathology report analysis: the traditional study approach and the more recent NLP framework approach (**Fig. 1**). This study compares these methodologies, highlighting their key differences and implications for future research.

The traditional study approach, often referred to as the AS-IS method, primarily encompasses manual structuring and analyzing semi-structured data. In the core of this method, researchers manually review semi-structured reports, which involves careful

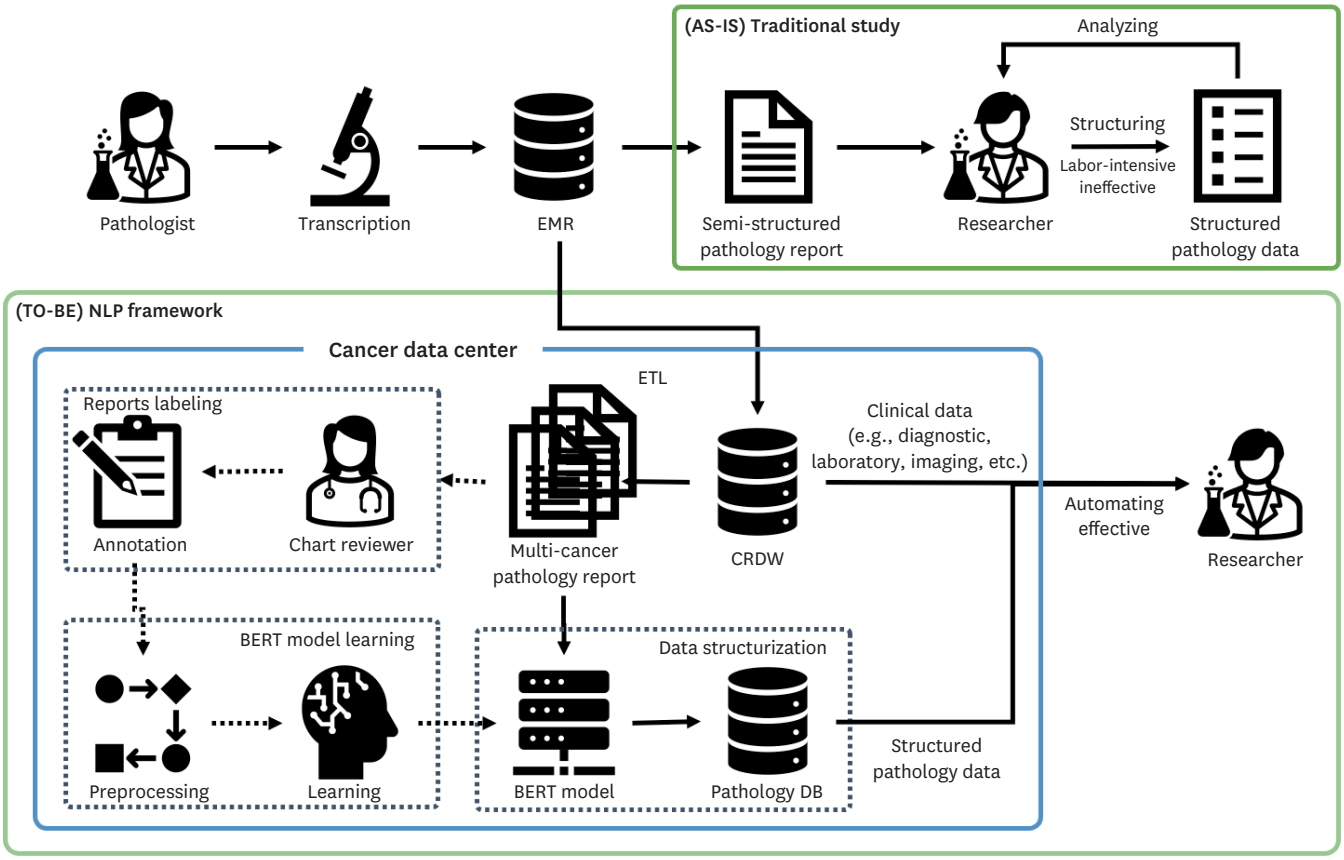

Fig. 1. Overview of the NLP framework.
NLP = natural language processing, DB = database, EMR = electronic medical record, CRDW = Clinical Research Data Warehouse, BERT = bidirectional encoder representations from transformers, ETL = extract transform load.

https://doi.org/10.3346/jkms.2026.41.e79

reading, interpretation, and structuring of data into a format suitable for specific research needs. Researchers meticulously organize information and often create structured datasets from semi-structured reports. This manual approach, although thorough, can be time-consuming and potentially lead to human error.

In contrast, the NLP framework approach, known as the TO-BE method, introduces a significant paradigm shift in the handling of pathology reports. This approach is characterized by the introduction of an NLP framework to automatically process and structure semi-structured data into a database suitable for analysis. This typically involves extracting reports from a Clinical Research Data Warehouse (CRDW), followed by the application of an NLP model such as BERT. These models are trained on a subset of manually annotated reports, allowing them to learn patterns and structures within pathology reports.

Data preprocessing

The data flowchart is shown in **Fig. 2**. All electronically available pathology reports from the National Cancer Center were stored as tables in the CRDW database, which contains medical records of patients with cancer. To ensure patient privacy, the CRDW assigns a unique anonymous identification key to each patient across all tables using a de-identification system.[18] We retrieved a total of 691,808 electronically available surgical pathology reports from the Department of Pathology at the National Cancer Center between January 1, 2005, and December 31, 2020. We extracted 964, 839, 1,008, and 1,215 pathology reports regarding liver, colorectal, stomach and breast cancers, respectively. After excluding 10% of the variables that lack value, we used the remaining 22, 24, 15, and 24 variables associated with liver, colorectal, stomach, and breast cancers, respectively, to create annotated data and develop an extraction algorithm.

To develop question-and-answer algorithms, it is necessary to obtain reference standards (i.e., an annotated corpus) through a review of surgical pathology reports for cancer and consultations with pathologists. Annotation guidelines were developed through repeated discussions with pathologists to establish standard annotations. To reduce errors that may occur between annotators, pathology reports were manually annotated by one annotator (a health information manager with a background in pathology) using an open-source label studio (HumanSignal, ver.1.1.0). The annotation results were finally reviewed by pathologists.

Statistical analysis

The BERT model was initially trained using Books Corpus and Wikipedia. However, it is important to note that the BERT model has some limitations regarding learning and understanding professional terms, especially in specialized fields such as medicine. To overcome this, BioBERT was developed and pre-trained specifically in technical terms, including medical terminology. BioBERT underwent transfer learning using a large dataset of nearly 18 billion words extracted from PubMed abstracts, making it highly proficient in medical languages. In addition, ClinicalBERT is an accurate language model that captures physician-assessed semantic relationships in clinical texts.

One of the most impressive capabilities of BERT-based models is their ability to be fine-tuned for specific tasks using annotated data. Herein, the task involved extracting breast cancer phenotypes from text. This task was approached in a question-and-answer format in which the model was trained to identify and answer questions related to 20 different types of variables associated with breast cancer. To achieve this, a carefully

https://doi.org/10.3346/jkms.2026.41.e79

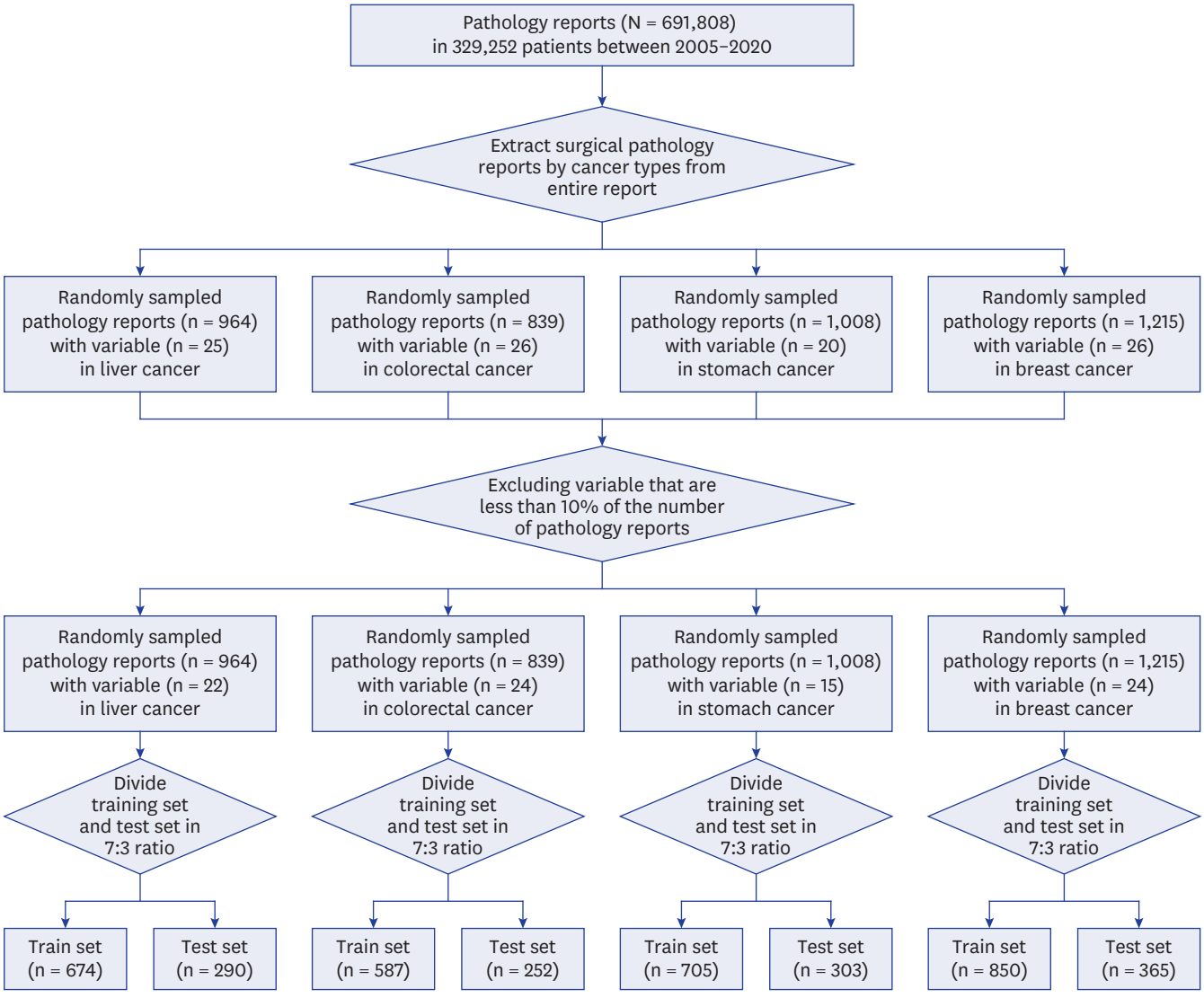

Fig. 2. Data flowchart of the study.

annotated training set was used to fine-tune the BERT-based model and ensure its effectiveness in this specific domain.

The key hyperparameters used in this study were set as follows. The maximum sequence length, determining the length of input text the model could process, was set to 128. The training batch size was 8, and the number of training epochs was 5. These hyperparameters were selected based on available computing resources to ensure efficient model training. These settings were consistently applied across the BERT, LSTM, and CNN models.

Like the BERT model, the LSTM model was trained using the same dataset and hyperparameters. However, it was specifically chosen to effectively learn temporal dependencies in the data while maintaining a manageable computational load. The LSTM model was used to capture long-term dependencies in text, which are critical for understanding the context and semantics of pathology reports.

https://doi.org/10.3346/jkms.2026.41.e79

Unlike the LSTM model, which focuses on sequential dependencies, the 1-dimensional-CNN model processes text by capturing spatial hierarchies of features. By applying the same hyperparameters, the CNN model was able to detect both local and global patterns in pathology reports, which are crucial for understanding the syntactic and semantic features of text data.

Ethics statement

This study was approved by the Institutional Review Board of National Cancer Center in Korea and the requirement for informed consent was waived (NCC2022-0105). In the pathology reports, data have been accessed since 11/8/2022.

RESULTS

Performance by artificial intelligence models

Table 1 presents a detailed comparison of the accuracy of 5 different NLP models (LSTM, CNN, BERT, BioBERT, and ClinicalBERT) in classifying various variables related to pathology reports.

The variable range was "Organ (OG)," "Operation name (ON)," and "Histology type (HT)." These variables are critical for understanding the specifics of a pathology report, and the ability of an NLP model to accurately classify them is crucial for automating the extraction of information from such reports.

Table 1. F1-score of five natural language processing models across multiple cancer types with 5-fold cross-validation results

Variables	BERT	BioBERT	ClinicalBERT	CNN	LSTM
Breast					
1-fold	0.9919	0.9893	0.9888	0.8729	0.9488
2-fold	0.9871	0.9894	0.9896	0.8178	0.9108
3-fold	0.9781	0.9778	0.9781	0.8612	0.9449
4-fold	0.9904	0.9976	0.9889	0.8213	0.9373
5-fold	0.9907	0.9840	0.9939	0.8555	0.9430
Average	0.9876	0.9876	**0.9879**	0.8457	0.9370
Colorectal					
1-fold	0.9696	0.9754	0.9637	0.8181	0.9452
2-fold	0.9861	0.9824	0.9765	0.8281	0.9326
3-fold	0.9877	0.9892	0.9879	0.7562	0.9425
4-fold	0.9932	0.9957	0.9880	0.8022	0.9214
5-fold	0.9728	0.9733	0.9755	0.8209	0.9284
Average	0.9819	**0.9832**	0.9783	0.8051	0.9340
Liver					
1-fold	0.9718	0.9656	0.9668	0.6575	0.7658
2-fold	0.9520	0.9483	0.9587	0.5782	0.7743
3-fold	0.9813	0.9910	0.9875	0.6635	0.7360
4-fold	0.9262	0.9442	0.9585	0.6675	0.7758
5-fold	0.9690	0.9578	0.9545	0.6291	0.7600
Average	0.9601	0.9614	**0.9652**	0.6392	0.7624
Stomach					
1-fold	0.9755	0.9302	0.9714	0.8886	0.9432
2-fold	0.9639	0.9513	0.9527	0.8899	0.9421
3-fold	0.9759	0.9221	0.9675	0.8955	0.9101
4-fold	0.9661	0.9512	0.9405	0.8614	0.9455
5-fold	0.9331	0.9675	0.9697	0.8032	0.9282
Average	**0.9629**	0.9445	0.9604	0.8677	0.9338

Bold values indicate the highest performance among the evaluated models.
BERT = bidirectional encoder representations from transformers, CNN = convolutional neural network, LSTM = long short-term memory.

https://doi.org/10.3346/jkms.2026.41.e79

Using 5-fold cross-validation, F1 scores were computed for each model across multiple cancer types. Statistical significance was assessed using McNemar's test. ClinicalBERT demonstrated significantly superior performance compared to LSTM and CNN models ($P < 0.05$) in most tasks. In contrast, no statistically significant differences were observed between ClinicalBERT and other BERT-based models, including BERT and BioBERT. Nevertheless, ClinicalBERT achieved the highest F1 scores in two out of four cancer types. Detailed McNemar's test results are provided in **Supplementary Table 1**.

The observed performance differences between traditional deep learning models (CNN, LSTM) and transformer-based models (BERT, BioBERT, ClinicalBERT) can be attributed to their fundamental architectural differences. While CNNs excel at capturing local patterns and LSTMs are designed to handle sequential data, they have inherent limitations when processing complex medical text. CNNs operate on fixed-size windows, potentially missing long-range dependencies crucial for understanding medical context. Similarly, although LSTMs can theoretically capture long-range dependencies, they often struggle with maintaining context over very long sequences common in pathology reports.

In contrast, transformer-based models utilize self-attention mechanisms that can directly model relationships between any words in the text, regardless of their distance. This architecture is particularly advantageous for pathology reports where key information may be scattered throughout the document and require understanding of complex medical terminology and context. ClinicalBERT's superior performance (F1-score 0.95) can be specifically attributed to its pre-training on clinical texts, which provides it with domain-specific knowledge that generalist models lack. All models were trained under a consistent 5-fold cross-validation framework. Hyperparameters were optimized through grid search or selected based on previous literature. Error analysis revealed that CNN and LSTM models particularly struggled with variables requiring long-range contextual understanding, such as distant metastasis or tumor border classification.

Among the BERT-based models, ClinicalBERT exhibited the best performance. Its architecture, which incorporates pre-training on both general and biomedical domain corpora, provides a distinct advantage in comprehending and categorizing medical terminology and concepts. This specialized training enabled ClinicalBERT to interpret the nuances of medical languages more accurately than the standard BERT model or BioBERT in numerous instances.

These findings suggest that for classifying variables from pathology reports, ClinicalBERT is the most effective and reliable model. Its superior performance across almost all variables indicates its robustness and reliability in automating information extraction from pathology reports. Conversely, the lower performances of the LSTM and CNN models suggest the need for further improvements.

Evaluation outcomes

Tables 2 and **3** present detailed analyses of performances of BERT-based models in classifying various variables related to stomach, liver, colorectal, and breast cancers. For each type of cancer, model performance was evaluated in terms of accuracy, recall, precision, and F1 scores, which are standard metrics used to measure the effectiveness of a classification model.

Our analysis of model performance across different cancer types revealed varying levels of accuracy. For stomach cancer, the model achieved perfect performance (accuracy = 1.0) across multiple variables, including angiolymphatic invasion, and operation name. However,

https://doi.org/10.3346/jkms.2026.41.e79

Table 2. Performance of the ClinicalBERT model according to pathology report variables for stomach and liver cancers

Variables	Accuracy	Recall	Precision	F1
Stomach				
Angiolymphatic invasion	1	1	1	1
Depth of invasion	0.8150	0.9803	0.8150	0.8607
Depth of invasion mucosa	0.6667	0.6667	0.6667	0.6667
Depth of invasion submucosa	0.8939	0.8990	0.8939	0.8929
Distant metastasis	0.3333	0.5000	0.3333	0.3889
Histologic type	0.8667	0.8396	0.8667	0.8365
Gross type	0.9960	0.9961	0.9960	0.9958
Tumor site	0.9684	0.9697	0.9684	0.9671
Lymphatic invasion	0.9960	0.9968	0.9960	0.9962
Lymph node	0.9951	0.9909	0.9951	0.9928
Operation name	1	1	1	1
Organ	0.9882	0.9765	0.9882	0.9823
Perineural invasion	0.9951	1	0.9951	0.9975
Surgical margin	0.9916	0.9916	0.9916	0.9916
Submucous fibrosis	0.9352	0.9129	0.9352	0.9229
Tumor border	0.9904	1	0.9904	0.9951
Venous invasion	0.9957	0.9916	0.9957	0.9935
Liver				
Additional chronic hepatitis	0.8409	0.8636	0.8409	0.8520
Additional cirrhosis	1	1	1	1
Additional dysplasia	0.9817	1	0.9817	0.9893
Bile duct invasion	1	1	1	1
Diagnosis	0.9859	0.9789	0.9859	0.9824
E&S major	0.9350	0.9351	0.9350	0.9348
E&S worst	0.9680	0.9760	0.9680	0.9719
Fatty change	0.9923	1	0.9923	0.9956
Gross type	0.9925	1	0.9925	0.9962
Histologic type	0.9846	0.9846	0.9846	0.9846
Hepatic vein invasion	1	1	1	1
Intrahepatic metastasis	0.9921	0.9842	0.9842	0.9881
Tumor site	0.9030	0.9120	0.9030	0.9066
Lymphatic invasion	1	1	1	1
Lymph node	0.8750	0.8125	0.8750	0.8333
Operation name	0.9577	0.9567	0.9577	0.9546
Organ	0.9931	1	0.9931	0.9965
Portal vein invasion	0.9921	0.9922	0.9921	0.9917
Surgical margin	0.9699	0.9699	0.9699	0.9699
Satellite nodule	1	1	1	1
Serosal invasion	1	1	1	1
Tumor size	0.9930	0.9894	0.9930	0.9906
Vascular invasion	0.9847	0.9924	0.9847	0.9885

BERT = bidirectional encoder representations from transformers, E&S = Edmondson and Steiner's histologic grade.

relatively lower performance was observed for distant metastasis (accuracy = 0.3333, F1 = 0.3889). This lower performance can be attributed to the limited number of metastasis cases in the pathology reports, with only 6 cases in the test set of which only a portion were correctly classified. In the case of liver cancer, excellent performance was noted for variables such as additional cirrhosis, bile duct invasion, hepatic vein invasion, lymphatic invasion, satellite nodule, serosal invasion, fatty change, gross type, organ, and tumor size with perfect or near-perfect scores (F1 > 0.99). The model consistently maintained a high performance across most variables, with no notably low scores observed in the updated results.

Regarding colorectal cancer, the model excelled in several important areas. Variables including Dworak's grade, extent of invasion, histologic grade, lymph node, lymphocytic response, operation name, pathology stage, peritonealization, preexist polyp, TME, tumor

https://doi.org/10.3346/jkms.2026.41.e79

Table 3. Performance of the BioBERT model according to pathology report variables for colorectal and breast cancers

Variables	Accuracy	Recall	Precision	F1
Colorectal				
Angiolymphatic invasion	0.9793	1	0.9793	0.9891
Dworak's grade	1	1	1	1
Extent of invasion	1	1	1	1
Histologic grade	1	1	1	1
Histologic type	0.9933	0.9933	0.9933	0.9933
Tumor site	0.9437	0.9671	0.9437	0.9502
Lymph node	1	1	1	1
Lymphocytic response	1	1	1	1
Operation name	1	1	1	1
Organ	0.9679	0.9576	0.9679	0.9617
Pathology stage	1	1	1	1
Perineural invasion	0.9521	0.9726	0.9521	0.9621
Peritonealization	1	1	1	1
Preexist polyp	1	1	1	1
Surgical margin	0.8889	1	0.8889	0.9259
TME	1	1	1	1
Tumor border	0.4730	1	0.4730	0.6156
Tumor size	1	1	1	1
Tumor budding	1	1	1	1
Tumor configuration	1	1	1	1
Tumor deposit	0.9714	1	0.9714	0.9810
Tumor perforation	0.9911	0.9822	0.9911	0.9866
Venous invasion	1	1	1	1
Breast				
Architectural pattern	0.9914	0.9885	0.9914	0.9894
Arteriovenous invasion	0.9836	1	0.9836	0.9917
Deep margin	0.9396	0.9868	0.9396	0.9558
Extensive intraductal	0.9714	0.9962	0.9714	0.9830
Histologic grade	0.9880	0.9960	0.9880	0.9913
Histologic type	0.9452	0.9453	0.9452	0.9442
Inferior margin	0.9784	0.9856	0.9784	0.9805
Intraductal component	0.9959	1	0.9959	0.9979
Lateral margin	0.9402	0.9670	0.9402	0.9480
Tumor site	0.9773	0.9646	0.9773	0.9709
Lymph node	0.9962	0.9962	0.9962	0.9962
Lymphovascular invasion	0.9835	1	0.9835	0.9917
Medial margin	0.9402	0.9655	0.9402	0.9505
Microcalcification	0.9866	1	0.9866	0.9932
Necrosis	0.9784	0.9957	0.9784	0.9868
Nipple margin	1	1	1	1
Nuclear grade	0.9914	0.9871	0.9914	0.9892
Operation name	0.9968	0.9968	0.9968	0.9966
Organ	1	1	1	1
Pathology stage	0.9964	1	0.9964	0.9980
Surgical margin	0.9856	0.9928	0.9856	0.9889
skin and nipple	0.9474	0.9660	0.9474	0.9534
Superficial margin	1	1	1	1
Superior margin	0.9960	0.9961	0.9960	0.9954
Tumor border	0.9837	0.9809	0.9837	0.9822
Tumor size	0.9924	0.9948	0.9900	0.9923

BERT = bidirectional encoder representations from transformers, TME = total mesorectal excision.

size, tumor budding, tumor configuration, and venous invasion all achieved perfect scores (accuracy = 1.0, F1 = 1.0). Performance was lower for tumor border classification (accuracy = 0.4730, F1 = 0.6156), primarily because while the model accurately predicts common outcomes like 'infiltration' or 'pushing,' it occasionally generates lengthy predictions (over 100 characters) for unusual cases, reducing overall accuracy.

https://doi.org/10.3346/jkms.2026.41.e79

The model also performed well in breast cancer classification, achieving accuracy for nipple margin, organ, and superficial margin (accuracy = 1.0, F1 = 1.0). Most other variables showed high performance, with F1 scores exceeding 0.95. Slightly lower yet still strong performance was observed for histologic type, lateral and medial margins (F1 = 0.9480 and 0.9505, respectively).

DISCUSSION

The ClinicalBERT-based NLP framework demonstrated remarkable effectiveness in extracting and analyzing data from pathology reports of various cancer types, including stomach, liver, colorectal, and breast cancers. The model achieved excellent performance metrics for numerous variables, with many exhibiting perfect or near-perfect scores (F1 > 0.95). In stomach cancer, F1 scores (F1 = 1.0) were achieved for clinically relevant variables such as tumor size, lymph node status, and operation name, which are essential for accurate staging and surgical planning. However, a lower performance was observed for distant metastasis (F1 = 0.3889), primarily due to the limited number of positive cases in the dataset. This limitation may result in missed metastasis cases and thus inappropriate staging or treatment decisions if the structured data is used uncritically.

In liver cancer, consistently high F1 scores (> 0.95) were observed across key variables such as cirrhosis, bile duct invasion, and tumor size, indicating that the model can accurately capture complex hepatic features from pathology narratives. For colorectal cancer, F1 scores (F1 = 1.0) were observed for many variables including pathology stage, lymphocytic response, and histologic grade. However, relatively lower performance was noted for tumor border (F1 = 0.6156), which may impact studies focusing on surgical margin characteristics or tumor behavior classification.

For breast cancer, the model also showed strong performance (F1 > 0.95) for most variables, with slightly lower scores for lateral and medial margins (F1 = 0.9480 and 0.9505, respectively). Although these scores are still high, subtle misclassifications in margin status may affect decisions related to re-excision or postoperative therapy in clinical settings.

The analysis revealed that variables with limited training samples or outlier values demonstrated lower accuracy levels, potentially leading to unreliable conclusions if researchers rely solely on the structured data output. To address this concern, we recommend expanding the training dataset specifically for these underperforming variables to enhance model accuracy and reliability. Additionally, by improving model performance and achieving high accuracy in initial structuring, the burden on researchers to manually validate or re-structure the extracted data can be substantially reduced.

The core of the NLP framework is its ability to automate extraction and structuring of semi-structured pathology reports. Once the NLP model is trained, it can be applied to large volumes of reports to rapidly convert them into a structured format. This automated process results in a dedicated pathology database in which structured data are stored in a readily available form suitable for analysis. This approach significantly reduces the manual effort required for data structuring and allows processing of much larger datasets than that feasible with manual methods.

https://doi.org/10.3346/jkms.2026.41.e79

Our study demonstrated several key advantages of the NLP framework for analyzing pathology reports. First, the NLP framework allows for the extraction of diverse variables from pathology records. Unlike traditional methods that rely on regular expressions for data parsing, the NLP model employed in our study enables the extraction of a wide range of information, including pathology grade, lesion location, treatment method, biomarker information, and immunopathology information.[3,4,6] This flexibility in variable extraction enhances the comprehensiveness and granularity of the structured data. Moreover, this approach enables researchers to access high-quality data by linking structured data and clinical information with the CRDW, further enhancing the value and utility of the extracted information. Second, the NLP framework contributes to the standardization of pathology records across different cancer types. NLP models can transform unstructured pathology reports into a structured format, ensuring consistency and uniformity in data representation. This standardization facilitates data integration, comparison, and analysis across different types of cancer, thus enabling more comprehensive research and insights. Moreover, the NLP framework is scalable and adaptable. Once the model is trained and validated, it can be applied to process pathology records of various cancer types, allowing for efficient and consistent data parsing. This scalability is particularly beneficial in the context of large-scale studies and multicenter collaborations, where a vast amount of data must be processed and analyzed.

Previous studies have mainly focused on extracting necessary variables from pathological records using NLP. However, the number of imported variables in these studies was limited. For instance, as shown in **Supplementary Table 2**, some studies only extracted a few variables, such as pathology grade, lesion location, and treatment method, from pathology records.[15,17,19-22] Other studies have focused on specific purposes, such as extracting biomarker information through a lymphoma classification tool or immunopathology information.[3] These studies have mainly been conducted with the aim of achieving specific objectives.

In addition, several studies have contributed to the field of NLP in the context of medical records. A previous study utilized NLP to extract trends in breast cancer molecular subtypes and Ki67 expression in South African women using rule-based information extracted from free-text pathology reports.[3] Another study focused on pattern-based information extraction from pathology reports for cancer registration,[4] aiming to improve the quality and efficiency of cancer registration through the use of NLP techniques. The field has evolved to incorporate more advanced techniques, with studies utilizing transfer learning approaches for biomedical named entity recognition using neural networks, and the development of specialized models such as GRAM-CNN that utilize local context information.[23] While these previous studies demonstrated the potential of NLP techniques in processing specific aspects of medical texts, our study extends this capability by developing a comprehensive framework that covers all relevant variables across multiple cancer types, enabling broader applications in pathology report analysis.

Despite these advantages, the limitations of this study should be acknowledged. Our model was developed based on annotated pathology reports from a single institution, which may limit its generalizability to other institutions. Our model was developed and validated using annotated pathology reports exclusively from the National Cancer Center, Korea, which may limit its generalizability to other institutions with different reporting formats, terminologies, or clinical practices. External validation studies are necessary. These studies should involve multiple institutions and diverse geographical regions. This will help ensure the broader applicability and robustness of our NLP framework. Moving forward, we intend to enhance the model's generalization capabilities through strategic fine-tuning using datasets from

https://doi.org/10.3346/jkms.2026.41.e79

multiple institutions. Future multi-center studies should evaluate the model's performance across different institutional settings, reporting styles, and patient populations. Additionally, the accuracy of the model in parsing pathological records of cancer types other than breast cancer may be limited.

Our study underscores the potential of NLP models, specifically ClinicalBERT, for effectively structuring data from pathology reports. By transforming unstructured text-heavy pathology reports into structured data, our NLP model can significantly enhance the potential of big data analysis in cancer research. This is of immense value as it allows researchers and healthcare professionals to harness the power of big data to advance cancer research and improve patient care. However, the varying performance of the model across different cancer types and variables underscores the need for further research and refinement.

SUPPLEMENTARY MATERIALS

Supplementary Table 1

McNemar test results for pairwise model comparisons across cancer type

Supplementary Table 2

Comparison of natural language processing approaches for data extraction from pathology reports

REFERENCES

1. Gardner J, Xiong L. An integrated framework for de-identifying unstructured medical data. *Data Knowl Eng* 2009;68(12):1441-51. CROSSREF
2. Ministry of Health and Welfare. *Health and Medical Data Utilization Guidelines*. Sejong, Korea: Ministry of Health and Welfare; 2020.
3. Achilonu OJ, Singh E, Nimako G, Eijkemans RMJC, Musenge E. Rule-based information extraction from free-text pathology reports reveals trends in South african female breast cancer molecular subtypes and Ki67 expression. *BioMed Res Int* 2022;2022(1):6157861. PUBMED | CROSSREF
4. Napolitano G, Fox C, Middleton R, Connolly D. Pattern-based information extraction from pathology reports for cancer registration. *Cancer Causes Control* 2010;21(11):1887-94. PUBMED | CROSSREF
5. Chang KP, Chu YW, Wang J. Analysis of hormone receptor status in primary and recurrent breast cancer via data mining pathology reports. *Open Med (Wars)* 2019;14(1):91-8. PUBMED | CROSSREF
6. Schadow G, McDonald CJ. *Extracting Structured Information From Free Text Pathology Reports*. Washington, D.C., USA: American Medical Informatics Association; 2003, 584.
7. Mohammad AH, Alwada'n T, Al-Momani O. Arabic text categorization using support vector machine, Naïve Bayes and neural network. *GSTF Int J Comput* 2016;5(1):16. CROSSREF
8. Pranckevičius T, Marcinkevičius V. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Balt J Mod Comput* 2017;5(2):221-32. CROSSREF
9. Graves JE, Pralus A, Fornoni L, Oxenham AJ, Caclin A, Tillmann B. Short- and long-term memory for pitch and non-pitch contours: Insights from congenital amusia. *Brain Cogn* 2019;136:103614. PUBMED | CROSSREF
10. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE* 1998;86(11):2278-324. CROSSREF
11. Alawad M, Gao S, Qiu JX, Yoon HJ, Blair Christian J, Penberthy L, et al. Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. *J Am Med Inform Assoc* 2020;27(1):89-98. PUBMED | CROSSREF

12. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019; June 2-7, 2019; Minneapolis, MN, USA. Stroudsburg, PA, USA: Association for Computational Linguistics; 2019.
13. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;36(4):1234-40. PUBMED | CROSSREF
14. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, McDermott M. Publicly available clinical BERT embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop; Jun 7, 2019; Minneapolis, MN, USA. Stroudsburg, PA, USA: Association for Computational Linguistics; 2019, 72-8.
15. Kim Y, Lee JH, Choi S, Lee JM, Kim JH, Seok J, et al. Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records. *Sci Rep* 2020;10(1):20265. PUBMED | CROSSREF
16. Kefeli J, Berkowitz J, Acitores Cortina JM, Tsang KK, Tatonetti NP. Generalizable and automated classification of TNM stage from pathology reports with external validation. *Nat Commun* 2024;15(1):8916. PUBMED | CROSSREF
17. Buckley JM, Coopey SB, Sharko J, Polubriaginof F, Drohan B, Belli AK, et al. The feasibility of using natural language processing to extract clinical information from breast pathology reports. *J Pathol Inform* 2012;3(1):23. PUBMED | CROSSREF
18. Cha HS, Jung JM, Shin SY, Jang YM, Park P, Lee JW, et al. The Korea Cancer Big Data Platform (K-CBP) for cancer research. *Int J Environ Res Public Health* 2019;16(13):2290. PUBMED | CROSSREF
19. Lam H, Nguyen F, Wang X, Stock A, Lenskaya V, Kooshesh M, et al. An accessible, efficient, and accurate natural language processing method for extracting diagnostic data from pathology reports. *J Pathol Inform* 2022;13:100154. PUBMED | CROSSREF
20. Glaser AP, Jordan BJ, Cohen J, Desai A, Silberman P, Meeks JJ. Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing. *JCO Clin Cancer Inform* 2018;2(2):1-8. PUBMED | CROSSREF
21. Choi HS, Song JY, Shin KH, Chang JH, Jang BS. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. *Radiat Oncol J* 2023;41(3):209-16. PUBMED | CROSSREF
22. Oliwa T, Maron SB, Chase LM, Lomnicki S, Catenacci DVT, Furner B, et al. Obtaining knowledge in pathology reports through a natural language processing approach with classification, named-entity recognition, and relation-extraction heuristics. *JCO Clin Cancer Inform* 2019;3(3):1-8. PUBMED | CROSSREF
23. Zhu Q, Li X, Conesa A, Pereira C. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics* 2018;34(9):1547-54. PUBMED | CROSSREF