Automatic recognition of abdominal lymph nodes from clinical text

Yifan Peng[1,3,*], Sungwon Lee[2,*], Daniel Elton[2], Tommy Shen[2], Yu-xing Tang[2], Qingyu Chen[1], Shuai Wang[2], Yingying Zhu[2,4], Ronald M. Summers[2,†], Zhiyong Lu[1,†]

[1]National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD 20894; [2]Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences Department, NIH Clinical Center, Bethesda, MD 20892; [3]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY 10065; [4]Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019

Abstract

Lymph node status plays a pivotal role in the treatment of cancer. The extraction of lymph nodes from radiology text reports enables large-scale training of lymph node detection on MRI. In this work, we first propose an ontology of 41 types of abdominal lymph nodes with a hierarchical relationship. We then introduce an end-to-end approach based on the combination of rules and transformer-based methods to detect these abdominal lymph node mentions and classify their types from the MRI radiology reports. We demonstrate the superior performance of a model fine-tuned on MRI reports using BlueBERT, called MriBERT. We find that MriBERT outperforms the rule-based labeler (0.957 vs 0.644 in micro weighted F1-score) as well as other BERT-based variations (0.913 - 0.928). We make the code and MriBERT publicly available at https://github.com/ncbi-nlp/bluebert, with the hope that this method can facilitate the development of medical report annotators to produce labels from scratch at scale.

1 Introduction

Lymph nodes are organs of the lymphatic system that are present throughout the body. Their status plays a pivotal role in the staging and treatment of cancer (Amin et al., 2017). The development of deep learning (DL) for computer vision has led to increasing interest in applying DL-based AI to identify and segment lymph nodes and detect lymph nodes and detecting lymph node metastasis in imaging studies, such as Magnetic Resonance Imaging (MRI). Applications of machine learning to MRI not only contribute to improving diagnostic accuracy but also reduce the workload of radiologists and enable them to spend additional time on high-level decision-making tasks. However, DL algorithms need to be sufficiently trained and evaluated using large-scale data before clinical adoption. Unlike general computer vision tasks, medical image analysis currently does not have enough annotated data (comparable to ImageNet and MS COCO), which is mainly because the conventional methods for harvesting labels cannot be applied in the clinical domain, as it requires extensive clinical expertise and because of security and privacy issues. Therefore, there is an unmet need to construct a large-scale annotated dataset of lymph nodes to increase the generalizability and robustness of the DL algorithms.

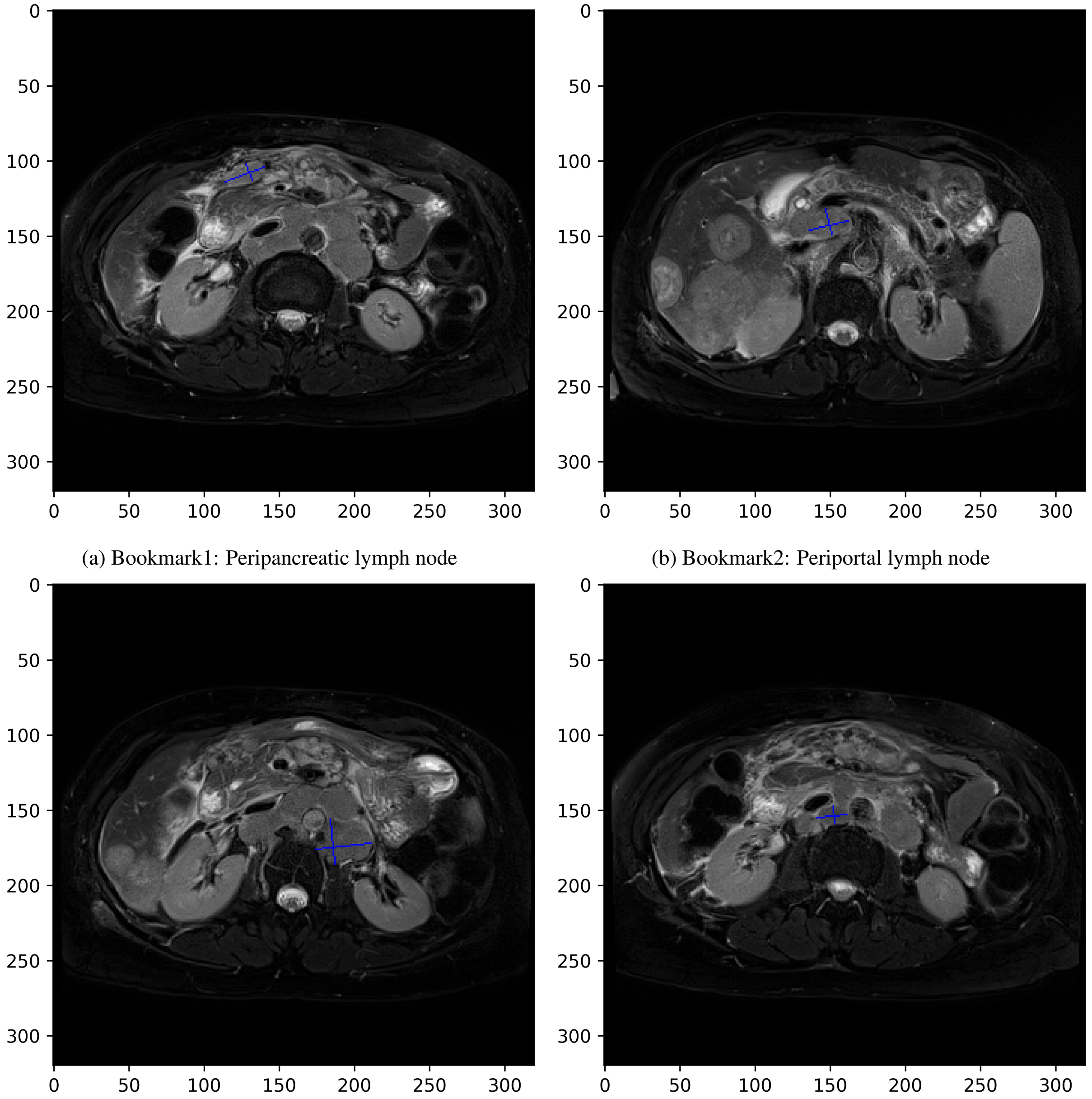
Radiologists report any abnormal lymph node detected in computed tomography (CT) and MRI exams by describing the regional name (type) of the lymph node. Example MRI scans and annotations are shown in Figure 1, where the radiologist describes the lymph node with the sentence "Abdominal/pelvic lymph nodes: There is intraperitoneal and retroperitoneal lymphadenopathy, for example, enlarged mesenteric/peripancreatic lymph node measuring **Bookmark1**[[(2.8 cm x 1.3 cm) (series 6, image 24)]], periportal lymph node measuring **Bookmark2**[[(2.8 cm x 1.7 cm) (series 6, image 19)]], retroperitoneal left paraaortic lymph node conglomerate measuring **Bookmark3**[[(3.8 cm x 3.1 cm) (series 6, image 22)]], and retroperitoneal aortocaval lymph node measuring **Bookmark4**[[(2.0 cm x 1.3 cm) (series 6, image 25)]]". The radiologist places a hyperlink (hereafter "bookmark") in the context to refer to the specified lymph node annotation in the image. Therefore, clinical reports provide a detailed and personalized account of assessments, offering a better context for clinical decision making and follow up.

Natural language processing (NLP) has been explored recently to unlock evidence buried in clin-

[*] These authors contributed equally to this work.

[†] Co-corresponding.

Proceedings of the 3rd Clinical Natural Language Processing Workshop, pages 101–110
November 19, 2020.



(a) Bookmark1: Peripancreatic lymph node

(b) Bookmark2: Periportal lymph node

(c) Bookmark3: Para-aortic lymph node

(d) Bookmark4: Interaortocaval lymph node

Figure 1: Sample sentence with lymph node bookmarks: “Abdominal/pelvic lymph nodes: There is intraperitoneal and retroperitoneal lymphadenopathy, for example enlarged mesenteric/peripancreatic lymph node measuring Bookmark1[[(2.8 cm x 1.3 cm) (series 6, image 24)]], periportal lymph node measuring Bookmark2[[(2.8 cm x 1.7 cm) (series 6, image 19)]], retroperitoneal left paraaortic lymph node conglomerate measuring Bookmark3[[(3.8 cm x 3.1 cm) (series 6, image 22)]], and retroperitoneal aortocaval lymph node measuring Bookmark4[[(2.0 cm x 1.3 cm) (series 6, image 25)]]. ”

ical narratives, making it available for large-scale analysis. In the clinical domain, NLP has been applied to identify positive, negative, and uncertain findings from radiology reports (Peng et al., 2018; Irvin et al., 2019; Yan et al., 2018). For MRI reports, NLP has been used to identify breast imaging lexicons for breast cancer (Sippo et al., 2013; Liu et al., 2019). However, most of these systems are rule-based, and few studies have investigated NLP in MRI reports of the lymph nodes.

To tackle these obstacles and challenges, this paper outlines a framework based on deep learning to harvest lymph node annotations and construct an annotated dataset of lymph nodes by automatically extracting lymph nodes from clinical reports. The contributions of this study are threefold: (1) We construct an ontology of 41 types of abdominal lymph nodes with a hierarchical relationship. (2) We develop a transformer-based deep learning module to extract and classify the abdominal lymph node types (or a non-abdominal lymph node or not a lymph node) for each bookmark mentioned in the sentence. (3) We make codes and pre-trained models publicly available.

The rest of the paper is organized as follows. We first present related work in Section 2. Then, we describe the method to construct the ontology and dataset in Section 3, followed by our experimental setup, results, and discussion in Section 4. We conclude with future work in the last section.

2 Related work

In recent years, there has been considerable interest in harvesting information and knowledge from free-text on electronic health records (EHRs) (Jensen et al., 2017). However, manually annotating a large dataset to fulfill the needs of deep learning models downstream is time-consuming and expensive. Therefore, researchers have applied NLP systems to identify structured labels from radiology reports (Irvin et al., 2019; Johnson et al., 2019; Wang et al., 2017; Smit et al., 2020).

Previous efforts in this area have focused mostly on two directions. One is the rule-based methods. NegEx, in combination with the Unified Medical Language System (UMLS), is a widely used algorithm that utilizes regular expressions to determine the negative concepts in the clinical narratives (Chapman et al., 2013; Aronson and Lang, 2010; Chapman et al., 2011). NegBio extended NegEx by utilizing universal dependencies and subgraph matching to detect both negative and uncertain lung diseases in chest X-rays and was used to generate labels for the NIH Chest X-ray and MIMIC-III-CXR datasets (Johnson et al., 2019; Wang et al., 2017; Peng et al., 2018). The CheXpert labeler further extended NegBio by increasing the rule sets and improving the NLP pipeline to construct report-level disease annotations (Irvin et al., 2019). CheXpert++ trained a hybrid rule- and BERT- based labeler on the radiograph domain but offers additional commentary on the utility of active-learning strategies to inform the interplay between the hybrid and rule-based labeler (McDermott et al., 2020).

The other direction is to apply machine learning methods to construct labels (Huang and Lowe, 2007; Clark et al., 2011; Xue et al., 2019; Peng et al., 2019a). Huang et al. described a hybrid approach to automatically detect negations in clinical radiology reports (Huang and Lowe, 2007). Clark et al. combine machine learning (conditional random field and maximum entropy) and rules to determine the assertion status of medical problems mentioned in clinical reports (Clark et al., 2011). Recently, deep learning approaches have also been studied intensively. Chen et al. applied CNNs to classify pulmonary embolism in chest CT reports (Chen et al., 2018). Drozdov et al. compared thirteen supervised classifiers and demonstrate that bidirectional long short-term memory (BiLSTM) networks with attention mechanisms effectively identify labels in CXR reports (Drozdov et al., 2020). Wood et al. present a transformer-based network for brain magnetic resonance imaging (MRI) radiology report classification, which automates this task by assigning image labels based on free-text expert radiology reports (Wood et al., 2020). Smit et al. introduced a BERT-based approach to medical image report labeling that exploits both the scale of available rule-based systems and the quality of expert annotations (Smit et al., 2020).

3 Methods

In this section, we first describe the process of constructing the abdominal lymph node ontology and gold-standard labels from the MRI reports associated with lymph nodes on MRI images. Then we demonstrate the development of the transformer-based method to detect lymph nodes from the reports.

3.1 Abdominal lymph node ontology construction

The labeling task in this study is to extract the presence of abdominal lymph nodes from radiology reports. Therefore, the first step is to construct the lymph node ontology. The challenge here is that the nomenclature of abdominal lymph nodes is complicated. Most of them are named after the anatomical organs their lymphatics are draining from, but some are named after an adjacent structure, and some are named for an anatomical compartment space. This makes them have confusing synonyms or sometimes overlapping areas, giving them a hierarchy. To make a standardized version of the abdominal lymph node ontology, we used three widely used guidelines (Amin et al., 2017) and textbooks (Harisinghani, 2013; Richter and Feyerabend, 2012) to establish the hierarchical relationship, representative synonyms, and relationships with overlapping areas.

3.2 MRI dataset

For model development and validation, we collected large-scale MRI studies from NIH Clinical Center, performed between Jan 2015 to Sept 2019, along with their associated radiology reports. (Figure 2). The majority (63%) of the MRI studies were from the oncology department. The initial search from the Picture Archiving and Communication System (PACS) database at the NIH Clinical Center returns 21,786 studies with 9,343 patients. We excluded non-abdomen studies and studies with missing reports. The final dataset consists of a total of 2,099 lymph node bookmarks from 1,379 studies of 917 unique patients, and their corresponding text reports retrospectively from the Picture Archiving and Communication System (PACS) database at the NIH Clinical Center. These lymph node labels were reviewed by a radiologist with 12 years of post-graduate experience. The study was a retrospective study and was approved by the Institutional Review Board with a waiver of informed consent. This data set comprised the reference (gold) standard for our evaluation and comparative analysis.

3.3 Framework

We developed a hybrid system to extract abdominal lymph nodes from the MRI reports. It consists of two modules: (1) a rule-based lymph node detection, and (2) a transformer-based lymph node classification (Figure 3).

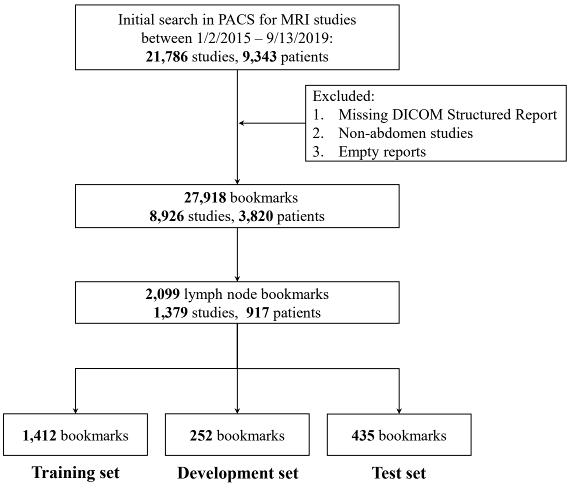

Initial search in PACS for MRI studies between 1/2/2015 – 9/13/2019: **21,786** studies, **9,343** patients
Excluded:
1. Missing DICOM Structured Report
2. Non-abdomen studies
3. Empty reports
27,918 bookmarks **8,926** studies, **3,820** patients
2,099 lymph node bookmarks **1,379** studies, **917** patients
1,412 bookmarks
252 bookmarks
435 bookmarks
Training set
Development set
Test set

Figure 2: The training, development, and test sets for classification of lymph node types from MRI reports.

3.3.1 Sentence extraction with potential lymph node bookmarks

In the reports of our institute, radiologists describe the lymph nodes and insert hyperlinks, size measurements, or slice numbers in the sentence to refer to the imaging findings of interest (called a bookmark). A bookmark thus is a hyperlink connection between the annotation in the image and the written description in the report. From the reports, we selected the full sentences that included the hyperlink, presuming that they had information most relevant to the connected image annotation.

In this step, we extract sentences with bookmarks that potentially link to lymph nodes. We first split the reports into sections. For our reports, the text is often organized into five sections: Clinical Indication, Technique, Comparison, Findings, and Impression. Among others, the "Findings" section lists the normal, abnormal, or potentially abnormal observations the radiologist saw in each area of the abdomen or pelvis in the exam. Hence, this section is often organized by organs such as the liver and kidney, blood vessels, and lymph nodes.

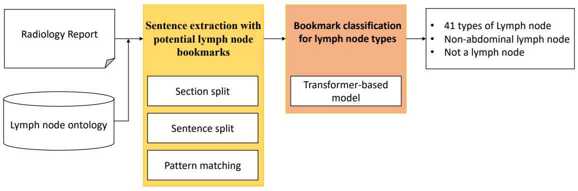


Figure 3: The architecture of the framework.

Table 1: The number of bookmarks in one sentence.

# bookmarks per sentence	n	(%)
1	1,457	69.4
2	409	19.5
3	149	7.1
4	63	3.0
5	15	0.7
6	6	0.3

Each section/subsection begins with a heading and ends with one or more empty lines. If available, the section headings were phrases from the beginning of a new line to a colon (e.g., "Liver and Gallbladder:"). We, therefore, use this information to split the reports into sections. Second, we tokenized the sentences using NLTK (Bird, 2006).

If a report contains the "Lymph node" subsection, we extracted sentences with lymph nodes from this subsection; otherwise, we extracted sentences with "lymph node" mentioned in the "Finding section" using regular expressions. We skipped the reports if it is not sectioned (0.3%). In our study, 85% of lymph node bookmarks are from the "Lymph node" subsection, and the remaining 16% are from reports with the "Lymph node" subsection but "Finding section" sections.

3.3.2 Bookmark classification for the abdominal lymph node type

After obtained candidate bookmarks that may link to lymph nodes, the next step is to classify bookmarks for the lymph node types. Here, we use the full sentences that included the bookmark, presuming that they had information most relevant to the connected image annotation. However, the bookmarked sentences often contain a complex mixture of information describing not only various bookmarked lymph nodes but also other bookmarked abnormalities. A sample sentence is shown in Figure 1. There are four bookmarks in a sentence, each of which has a different lymph node type. Table 1 shows that more than 30% of sentences have at least two bookmarks.

To solve this problem, we developed a transformer-based deep learning module with 43 labels (41 abdominal lymph node types, non-abdominal lymph node, and not a lymph node). Specifically, we treat the lymph node recognition task as a sentence classification by replacing the bookmark of interest in the sentence with a predefined tag BMK. Suppose that h_0 is the output embedding of the token [CLS], the probability that a bookmark labeled as class c is predicted by a fully connected layer and a logistic regression with softmax:$P(c|X) = softmax(ah_0 + b)$. We fine-tune the model on the training set using the categorical cross-entropy loss, $-\sum_c \delta(y_c = \hat{y}) log P(c|X)$ where $\delta(y_c = \hat{y}) = 1$ if the classification $\hat{y}$ of X is the correct ground-truth for the class $c \in C$; otherwise $\delta(y_c = \hat{y}) = 0$.

BERT is a contextualized word representation model that is pretrained based on a masked language modeling using bidirectional transformers (Devlin et al., 2019). In this paper, we fine-tuned the model using the BlueBERT base model (Peng et al., 2019b). The BlueBERT was pre-trained on the combination of PubMed and MIMIC-III clinical notes. We also compared the performance of our method using other BERT variants.

4 Results

4.1 Abdominal lymph node ontology

We construct an ontology of 41 abdominal lymph nodes relevant to MRI (Figure 4). Because of the nature of lymph node nomenclature, the labels had to have a hierarchical structure and some labels overlapped with others (Harisinghani, 2013; Richter and Feyerabend, 2012; Amin et al., 2017). Those subgroups include coarse, high-level lymph nodes such as "mediastinal lymph node", "retroperitoneal lymph node", and "pelvic lymph node", as well as fine-grained lymph nodes such as "perigastric lymph node along greater curvature" and "pericecal lymph node". Table 2 shows the distribution of lymph nodes in the dataset, which is imbalanced. The majority of abdominal lymph nodes in the dataset are periportal and para-aortic lymph nodes.

4.2 Results of the lymph node classification

We trained the model on one NVIDIA® V100 GPU using the TensorFlow framework26. We used the Adamax optimizer (Kingma and Ba, 2015) with a learning rate of 10^{-5} and a batch size of 32. We used the BlueBERT base model as the domain-specific language model. As a result, all the tokenized texts using wordpieces (Wu et al., 2016) were chopped to spans no longer than 128 tokens. We set the maximum number of epochs to 30.

To evaluate the performance of the framework, we use 70% for training, 10% for development, and

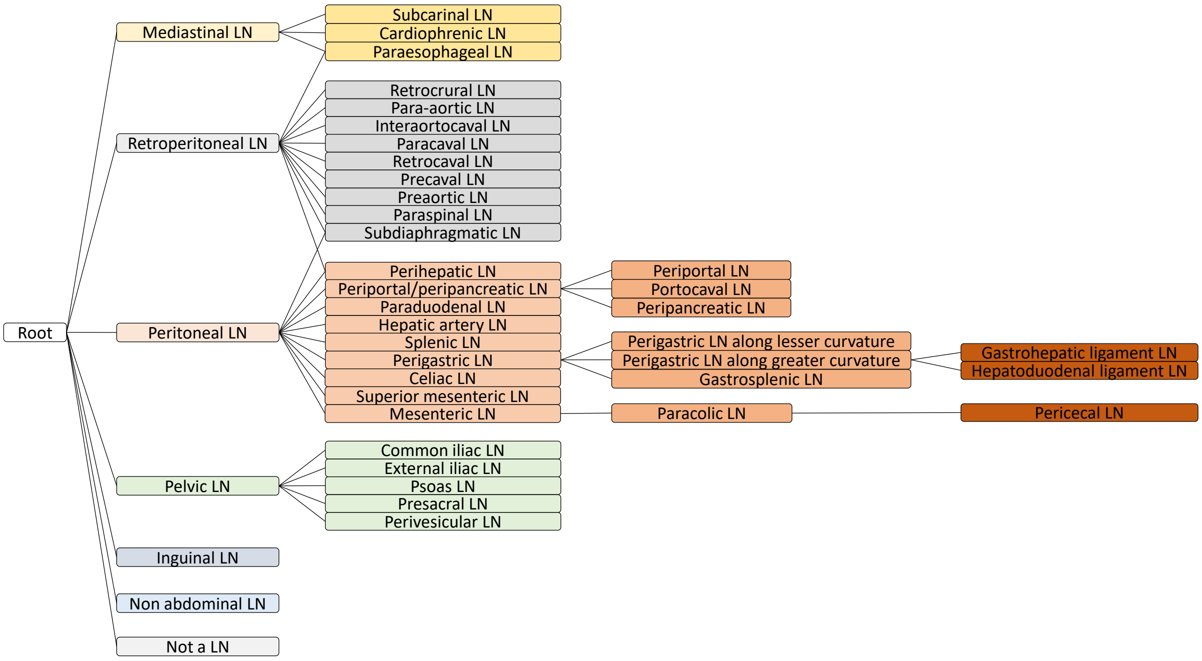


Figure 4: The abdominal lymph node (LN) ontology.

Table 2: The distribution of lymph node in the dataset.

Lymph node	n	(%)	Lymph node	n	(%)
Periportal	300	14.30	Paraduodenal	11	0.50
Para-aortic	278	13.20	Subcarinal	9	0.40
Retroperitoneal	257	12.20	Superior mesenteric	9	0.40
Mesenteric	186	8.90	Paraesophageal	8	0.40
Portocaval	125	6.00	Peritoneal	7	0.30
Peripancreatic	120	5.70	Paraspinal	7	0.30
Interaortocaval	95	4.50	Paracolic	7	0.30
Gastrohepatic ligament	73	3.50	Pericecal	7	0.30
Retrocrural	44	2.10	External iliac	6	0.30
Paracaval	39	1.90	Pelvic	6	0.30
Retrocaval	32	1.50	Inguinal	6	0.30
Mediastinal	26	1.20	Perigastric LN along GC	5	0.20
Periportal/peripancreatic	24	1.10	Perigastric	4	0.20
Common iliac	21	1.00	Hepatoduodenal ligament	4	0.20
Cardiophrenic	20	1.00	Hepatic artery	3	0.10
Precaval	19	0.90	Splenic	2	0.10
Psoas	18	0.90	Presacral	1	0.00
Celiac	17	0.80	Perigastric LN along LC	1	0.00
Perihepatic	14	0.70	Non-abdominal LN	238	11.30
Subdiaphragmatic	12	0.60	Not a LN	26	1.20
Preaortic	12	0.60			

GC - greater curvature. LC - lesser curvature

20% for testing. Table 3 shows the performance of our systems on the classification of 5 coarse-grained lymph node types by (P)recision, (R)ecall, and (F)1-score. The micro metrics count the total true positives, false negatives, and false positives across all lymph node types. The macro metrics calculate precision, recall, and F1 for each lymph node type and find their unweighted mean. The

Table 3: Test results on the classification of 5 coarse-grained lymph node types.

Lymph nodes	P	R	F
Mediastinal LN	0.778	1.000	0.875
Retroperitoneal LN	0.975	0.994	0.985
Peritoneal LN	0.959	0.989	0.974
Pelvic LN	1.000	0.923	0.960
Inguinal LN	1.000	1.000	1.000
Non-abdominal LN	0.952	0.784	0.860
Not a LN	1.000	0.500	0.667
micro	0.959	0.959	0.959
macro	0.952	0.884	0.903
micro weighted	0.960	0.959	0.957

weighted metrics calculate precision, recall, and F1 for each lymph node type and find their average weighted by the number of true instances for each type. Our system achieved an overall precision of 0.960, recall of 0.959, and F1-score of 0.957. We achieved F1-score ≥ 0.850 on all coarse-grained lymph node types. On the other hand, we observed that on "negative" cases (not a lymph node), the recall is 0.5. This is because the dataset has fewer negative instances (26) in total, which may not be sufficient to train and test the model. In the future, more negative cases shall be manually included to handle the imbalanced dataset. However, we consider it not a major issue in our framework since the first step utilizes rigid extraction patterns and achieves high precision.

Table 4 shows the performance on the classification of all fine-grained lymph node types. Our system achieved an overall precision of 0.925, recall of 0.913, and F1-score of 0.912. We achieved F1-score 1.00 on 8 types, ≥ 0.90 on 17 types, and ≥ 0.80 on 23 types.

We also compare our model on BERT variants: ClinicalBERT (Alsentzer et al., 2019), BioBERT (Lee et al., 2020), and BlueBERT. The ClinicalBERT was pretrained on MIMIC-III generic clinical text. The BioBERT was pretrained on PubMed. For reference, we include a rule-based system where the type of lymph node is selected based on the nearest keyword (e.g., cardiophrenic, inguinal, etc.) from the bookmark in the sentence. Table 5 shows that deep-learning-based methods can successfully classify the type of each lymph node mentioned in the sentences. The system using BlueBERT (MriBERT) outperforms that using BioBERT. This observation shows the impact of using clinical notes during the pre-training process. On the other hand, the system using ClinicalBERT achieved lower performance. It may suggest that the MIMIC-III clinical text alone may not be large enough to sufficiently pre-train the BERT model.

5 Conclusion

In this study, we introduced an ontology of 41 types of abdominal lymph nodes with a hierarchical relationship. We then proposed an end-to-end framework for combining rules and deep learning for accurate bookmark classification for lymph node types from MRI reports. In this framework, the rule-based method is first used to extract sentences with potential lymph node bookmarks. Then a BERT-based model pretrained on MRI reports was used to classify each bookmark into one of 41 types of abdominal lymph node, non-abdominal lymph nodes, or not a lymph node. We evaluated our framework on 2,099 bookmarks manually annotated by a radiological expert. We also compared our framework with a rule-based system and other BERT-based models. We find that our framework achieved 0.912 in F1-score, which outperforms the rule-based system and other BERT variations.

Our study has several limitations. First, our model is limited to the 41 abdominal lymph nodes. While we believe the list is comprehensive, we may miss some lymph node types due to training corpus bias. Second, our evaluation is performed on a single corpus. Cross-institutional experiments need to be performed in the future to evaluate the generalizability of the model.

While our work only scratches the surface of using text mining techniques and deep learning to extract the lymph node from radiology reports, we hope it will shed light on the development of generalizable NLP models that can extract highly accurate labels.

Acknowledgment

This work was supported by the Intramural Research Programs of the NIH National Library of Medicine and NIH Clinical Center. This work was also supported by the National Library of Medicine of the NIH under award number 4R00LM013001. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

Table 4: Test results on the classification of fine-grained lymph node types.

Lymph nodes	P	R	F	Lymph nodes	P	R	F
Periportal	0.903	0.933	0.918	Subdiaphragmatic	1.000	0.667	0.800
Para-aortic	0.902	0.982	0.940	Preaortic	0.500	0.333	0.400
Retroperitoneal	0.980	0.962	0.971	Paraduodenal	1.000	0.667	0.800
Mesenteric	0.900	0.947	0.923	Subcarinal	0.667	1.000	0.800
Portocaval	0.889	0.960	0.923	Superior mesenteric	1.000	0.500	0.667
Peripancreatic	0.923	1.000	0.960	Paraesophageal	1.000	1.000	1.000
Interaortocaval	1.000	1.000	1.000	Peritoneal	1.000	1.000	1.000
Gastrohepatic ligament	0.938	1.000	0.968	Paraspinal	1.000	1.000	1.000
Retrocrural	1.000	1.000	1.000	Paracolic	1.000	0.500	0.667
Paracaval	0.667	0.500	0.571	Pericecal	0.500	1.000	0.667
Retrocaval	0.667	0.857	0.750	External iliac	1.000	1.000	1.000
Mediastinal	0.625	0.833	0.714	Pelvic	1.000	1.000	1.000
Periportal/peripancreatic	0.833	1.000	0.909	Inguinal	0.667	1.000	0.800
Common iliac	1.000	1.000	1.000	Perigastric LN along LC	0.500	1.000	0.667
Cardiophrenic	0.750	0.750	0.750	Non-abdominal LN	0.975	0.765	0.857
Precaval	0.750	0.750	0.750	Not a LN	1.000	0.333	0.500
Psoas	1.000	1.000	1.000	*micro*	0.913	0.913	0.913
Celiac	1.000	1.000	1.000	*macro*	0.861	0.859	0.839
Perihepatic	1.000	0.667	0.800	*micro weighted*	0.925	0.913	0.912

GC - greater curvature. LC - lesser curvature

Table 5: Test results of various methods on lymph node classification.

Models	Coarse-grained LN types			Fine-grained LN types		
	P	R	F	P	R	F
Rule-based	0.827	0.579	0.644	0.699	0.453	0.533
ClinicalBERT	0.914	0.915	0.913	0.878	0.878	0.874
BioBERT	0.932	0.931	0.928	0.896	0.887	0.885
BlueBERT (MriBERT)	**0.960**	**0.959**	**0.957**	**0.925**	**0.913**	**0.912**

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mahul B. Amin, American Joint Committee on Cancer, and American Cancer Society, editors. 2017. *AJCC Cancer Staging Manual*, eight edition / editor-in-chief, mahul b. amin, md, fcap ; editors, stephen b. edge, md, facs [and 16 others] ; donna m. gress, rhit, ctr - technical editor ; laura r. meyer, capm - managing editor edition. American Joint Committee on Cancer, Springer, Chicago IL.

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–236.

Steven Bird. 2006. NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, pages 69–72. Association for Computational Linguistics.

Brian E. Chapman, Sean Lee, Hyunseok Peter Kang, and Wendy W. Chapman. 2011. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. *Journal of Biomedical Informatics*, 44(5):728–737.

Wendy W Chapman, Dieter Hillert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E Chapman, Mike Conway, Melissa Tharp, Danielle L Mowery, and Louise Deleger. 2013. Extending the NegEx lexicon for multiple languages. *Studies in health technology and informatics*, 192:677–681.

Matthew C. Chen, Robyn L. Ball, Lingyao Yang, Nathaniel Moradzadeh, Brian E. Chapman, David B. Larson, Curtis P. Langlotz, Timothy J. Amrhein, and Matthew P. Lungren. 2018. Deep Learning to Classify Radiology Free-Text Reports. *Radiology*, 286(3):845–852.

Cheryl Clark, John Aberdeen, Matt Coarr, David Tresner-Kirsch, Ben Wellner, Alexander Yeh, and Lynette Hirschman. 2011. MITRE system for clinical assertion status classification. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):563–567.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ignat Drozdov, Daniel Forbes, Benjamin Szubert, Mark Hall, Chris Carlin, and David J. Lowe. 2020. Supervised and unsupervised language modelling in Chest X-Ray radiological reports. *PLOS ONE*, 15(3):e0229963.

Mukesh G. Harisinghani, editor. 2013. *Atlas of Lymph Node Anatomy*. Springer, New York.

Yang Huang and Henry J. Lowe. 2007. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association*, 14(3):304–311.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, and Henrik Marklund. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.

Kasper Jensen, Cristina Soguero-Ruiz, Karl Oyvind Mikalsen, Rolv-Ole Lindsetmo, Irene Kouskoumvekaki, Mark Girolami, Stein Olav Skrovseth, and Knut Magne Augestad. 2017. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific Reports*, 7(1):46226.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, pages 1–15.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)*, 36(4):1234–1240.

Yi Liu, Li-Na Zhu, Qing Liu, Chao Han, Xiao-Dong Zhang, and Xiao-Ying Wang. 2019. Automatic extraction of imaging observation and assessment categories from breast magnetic resonance imaging reports with natural language processing. *Chinese Medical Journal*, 132(14):1673–1680.

Matthew B. A. McDermott, Tzu Ming Harry Hsu, Wei-Hung Weng, Marzyeh Ghassemi, and Peter Szolovits. 2020. CheXpert++: approximating the chexpert labeler for speed,differentiability, and probabilistic output. *arXiv:2006.15229 [cs, stat]*.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. NegBio: A high-performance tool for negation and uncertainty detection in radiology reports. In *AMIA Joint Summits on Translational Science Proceedings*.

AMIA Joint Summits on Translational Science, volume 2017, pages 188–196.

Yifan Peng, Ke Yan, Veit Sandfort, Ronald M. Summers, and Zhiyong Lu. 2019a. A self-attention based deep learning method for lesion attribute detection from CT reports. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019b. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 58–65.

E Richter and Thomas Feyerabend. 2012. *Normal Lymph Node Topography: CT Atlas*. Springer Berlin Heidelberg, Berlin.

Dorothy A. Sippo, Graham I. Warden, Katherine P. Andriole, Ronilda Lacson, Ichiro Ikuta, Robyn L. Birdwell, and Ramin Khorasani. 2013. Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing. *Journal of Digital Imaging*, 26(5):989–994.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. 2020. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. *arXiv:2004.09167 [cs]*.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471. IEEE.

David A. Wood, Jeremy Lynch, Sina Kafiabadi, Emily Guilhem, Aisha Al Busaidi, Antanas Montvila, Thomas Varsavsky, Juveria Siddiqui, Naveen Gadapa, Matthew Townend, Martin Kiik, Keena Patel, Gareth Barker, Sebastian Ourselin, James H. Cole, and Thomas C. Booth. 2020. Automated Labelling using an Attention model for Radiology reports of MRI scans (ALARM). *arXiv:2002.06588 [cs]*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Kui Xue, Yangming Zhou, Zhiyuan Ma, Tong Ruan, Huanhuan Zhang, and Ping He. 2019. Fine-tuning BERT for joint entity and relation extraction in Chinese medical text. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 892–897. IEEE.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M. Summers. 2018. DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging (Bellingham, Wash.)*, 5(3):036501.