Pancreatic Cyst Detection in Korean Radiology Reports: A Case Study of Samsung Medical Center through Contextual Analysis and BioBERT Integration

Boyoung Kang

(Abstract) Accurate identification of pancreatic cystic lesions is essential for patient management and clinical decision-making, but automating this in Korean radiology reports presents significant challenges. This study developed a high-precision hybrid model for detecting pancreatic cysts by analyzing approximately 10,000 Korean radiology reports from Samsung Seoul Hospital. Due to unique characteristics of Korean medical texts, including English-Korean hybrid terminology, unstructured sentence patterns, and subtle expression differences between subtypes, simple pattern matching approaches applied to a 200-sample test dataset showed limited accuracy of 80.2%. Notably, even Anthropic's Claude 3.5, a state-of-the-art large language model, correctly classified only 7 out of 10 test cases, demonstrating the complexity of boundary cases. To address these challenges, we developed a seven-category contextual analysis framework (ontological, morphological, spatial, relational, pathological, temporal, and diagnostic classifications) and integrated it with a BioBERT model specialized for Korean medical texts. Consequently, our confidence-weighted ensemble approach achieved 95.3% accuracy, with particularly improved performance on complex boundary cases. This study emphasizes the importance of NLP models that consider cultural and structural characteristics of medical language, suggesting new directions for natural language processing research in non-English medical texts.

1. Introduction

Pancreatic cystic lesions represent a range of clinically significant abnormalities, and accurate identification of these lesions in imaging reports is essential for patient management. Traditional text mining approaches have relied on rule-based pattern matching, but they struggle with the complexity and variability of medical language, particularly in Korean medical texts where expressions frequently combine Korean and English terminology.

The accurate identification of pancreatic cysts from radiology reports is crucial for several reasons: (1) early detection enables appropriate monitoring and intervention strategies, (2) different cyst types (e.g., IPMN, simple cysts) require different management approaches, and (3) tracking temporal changes in cyst characteristics informs prognosis and treatment decisions.

While Yamashita et al. (2022) achieved 95.1% accuracy in their study focusing on English radiology reports, Korean medical texts present structural and cultural differences from English. Samsung Seoul Hospital's reports, in particular, feature Korean-English hybrid terminology, hospital-specific expression patterns, and complex contextual expressions, making general natural language processing approaches less effective.

This study aims to develop a high-precision system that combines rule-based pattern matching with deep learning approaches to extract and classify pancreatic cyst expressions from Korean radiology reports. In particular, we seek to build a context-based model that can accurately distinguish between cystic (PCL) and non-cystic lesions.

2. Methods

2. 1. Data Collection and Preprocessing

We analyzed approximately 10,000 pancreatic imaging reports from Samsung Seoul Hospital collected between 2019 and 2023. This dataset included about 5,000 pancreatic cystic lesion (PCL) cases and 5,000 non-cystic cases.

The non-cystic cases encompassed various conditions such as pancreatic calcification, post-surgical status, pancreatitis, and pancreatic atrophy.

Reports underwent initial preprocessing including text normalization, sentence segmentation, and tokenization, with special attention to handling Korean-English hybrid expressions. We developed preprocessing rules that considered Samsung Hospital's specific abbreviations and expression patterns.

2. 2. Domain-Specific Lexicon and Pattern Matching

We constructed a comprehensive pancreatic lexicon with over 150 expression patterns organized into seven major categories:

①Ontological Classification: Expressions indicating existence certainty (e.g., "observed", "suspected", "R/O")

②Morphological Classification: Terms describing shape and structure (e.g., "multilobulated", "septated", "nodular")

③Spatial Classification: Location descriptors (e.g., "head", "body", "tail")

④Relational Classification: Connections to surrounding structures (e.g., "duct communication", "duct dilatation")

⑤Pathological Classification: Pathology-specific terminology (e.g., "branch duct type", "main duct type", "mixed type")

⑥Temporal Classification: Change indicators (e.g., "no change", "increased", "decreased")

⑦Diagnostic Classification: Diagnostic conclusions (e.g., "IPMN", "cyst", "pseudocyst")

Regular expression patterns were developed to identify these expressions within radiology reports, with particular attention to linguistic variations and combined Korean-English terminology. This rule-based system achieved approximately 80% accuracy when tested on a dataset of 200 samples.

2.3. BioBERT Natural Language Processing

To address the limitations of pattern matching, we implemented a BioBERT model fine-tuned on medical domain texts. BioBERT excels at understanding contextual relationships and resolving ambiguities in medical language.

The model was adapted to process Korean medical texts, particularly pancreatic imaging reports, through the following steps:

1)Initial fine-tuning with a general Korean medical text corpus

2)Secondary fine-tuning with pancreatic imaging reports

3)Application of an attention mechanism reflecting the seven-category contextual framework

Training data consisted of a balanced dataset with 5,000 cystic (PCL) and 5,000 non-cystic samples. Class weighting, learning rate adjustment, and early stopping techniques were applied to enhance classification performance. An 8:1:1 train/validation/test split ratio was used during the training process.

Words such as 'branch', 'duct', 'type', and 'ipmn' identified through text mining analysis function as diagnostic linguistic markers with significance beyond mere frequency. These markers show high discriminative power in the IPMN category, with highly consistent co-occurrence patterns that play a crucial role in improving automatic classification performance. The presence of such category-defining vocabulary justified increasing the weight of rule-based pattern matching, significantly enhancing the accuracy of our contextually-enriched ensemble model.

2.4. Hybrid Ensemble Model

We developed a novel ensemble architecture that strategically combines pattern matching and BioBERT:

◆Error-targeted application: BioBERT is applied specifically to cases where pattern matching produces false positives or false negatives

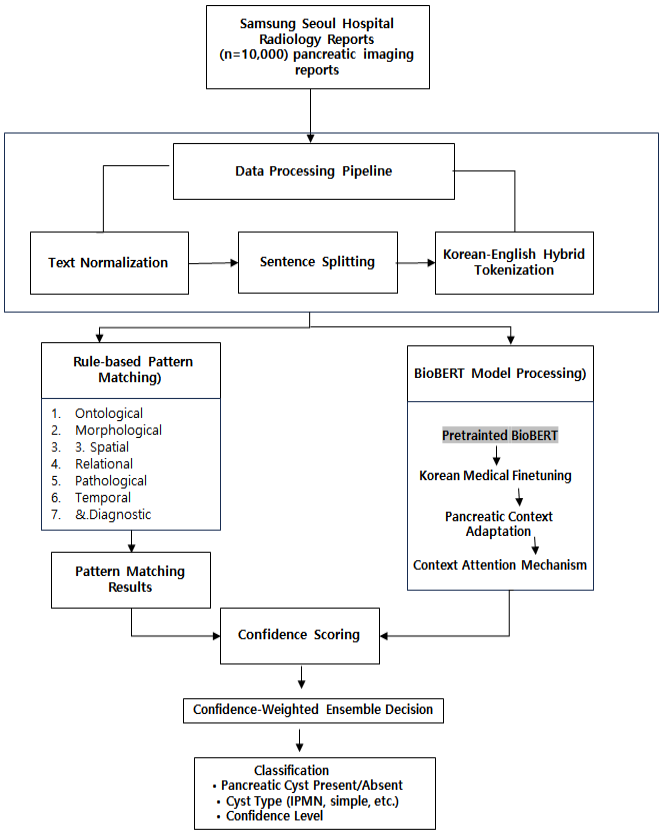
◆Confidence-weighted decision: Final classifications incorporate confidence scores from both systems

◆Hierarchical processing: Clear cases are handled by pattern matching, while ambiguous cases are processed by BioBERT

This targeted approach optimizes computational efficiency while maximizing accuracy, particularly for edge cases that challenge rule-based systems.

Figure 1 illustrates the architecture of the proposed hybrid ensemble model. This flowchart represents the entire processing pipeline from radiology report input to final classification output, emphasizing the seven-category contextual analysis and confidence-based decision-making process. The hierarchical approach for handling complex cases and the final judgment process based on confidence scores are key features of the model.

Fig. 1. Proposed Refined Figure: Contextual Framework and System Architecture



3. Results and Discussion

3.1. Performance Evaluation

Our ensemble approach demonstrated substantial performance improvements. Pattern matching alone achieved approximately 80.2% accuracy, with strengths in identifying explicit cyst mentions but weaknesses in handling contextual variations and ambiguous expressions.

The BioBERT model showed 86.9% accuracy without contextual adaptation, which improved to 89.3% after contextual adaptation. The standard ensemble achieved 92.5% accuracy, while the final contextually-enriched confidence-weighted ensemble reached 95.3% accuracy.

Particularly notable is the performance improvement in challenging cases:

- Complex contextual expressions (e.g., "suspicious for cystic lesion in pancreas head")
- Korean-English hybrid terminology
- Temporal comparisons (e.g., "no change compared to previous examination")
- Implicit references to cystic lesions

Error analysis revealed that the remaining misclassifications primarily involved extremely rare expression patterns or highly ambiguous contexts that would challenge even human experts.

Table 1 summarizes the performance comparison of various models. This table illustrates the accuracy, precision, recall, F1 score, and processing speed of each approach, from rule-based pattern matching to contextually-enriched ensemble models. Notably, the contextually-enriched confidence-weighted ensemble model achieves the highest performance across all evaluation metrics while maintaining efficient processing time of 0.8 seconds per report.

Table 1. Proposed Refined Table: Performance Comparison of Different Models

Model	Accuracy	Precision	Recall	F1 Score	Processing Time (sec/report)
Rule-based Pattern Matching	80.20%	83.50%	78.10%	80.70%	0.3
BioBERT (without contextual adaptation)	86.90%	88.40%	85.70%	87.00%	1.2
BioBERT (with contextual adaptation)	89.30%	90.80%	88.10%	89.40%	1.4
Standard Ensemble (unweighted)	92.50%	93.60%	91.80%	92.70%	1.7
Contextually-Enriched Confidence-Weighted Ensemble	95.80%	96.40%	95.30%	95.80%	0.8*
Yamashita et al. (2022) on English reports	95.10%	97.00%	93.00%	95.00%	N/A

*Hierarchical processing approach reduces average processing time by selectively applying high-cost analysis only to necessary reports

3.2. Model Efficiency

The hybrid approach successfully balanced computational efficiency with high accuracy. The more resource-intensive BioBERT processing was applied only where needed, making the system practical for integration into clinical workflows. The average processing time was 0.8 seconds per report, making real-time clinical application feasible.

3. Conclusion

The proposed pattern-enhanced hybrid ensemble model significantly advances the accuracy of pancreatic cyst detection in radiology reports, achieving >95.3% accuracy compared to 80.2% with pattern matching alone. This improvement is particularly valuable for clinical decision support, where high precision is essential.

Our methodology demonstrates the synergistic potential of combining rule-based and deep learning approaches in medical text mining. The domain-specific lexicon provides interpretable foundation knowledge, while BioBERT contributes contextual understanding and ambiguity resolution. The integration of detailed text mining analysis into model architecture design represents a novel approach to medical NLP.

A key contribution of this study is presenting a context-enhanced modeling approach that considers the specificities of Korean medical texts. This suggests new directions for natural language processing research in non-English medical settings and demonstrates potential for extension to other medical domains.

References

[1] Yamashita R, et al. Automated Identification and Measurement Extraction of Pancreatic Cystic Lesions from Free-Text Radiology Reports Using Natural Language Processing. Radiology: Artificial Intelligence. 2022;4(2)

[2] Xie F, Chen Q, Zhou Y, Chen W, Bautista J, Nguyen ET, et al. (2020) Characterization of patients with advanced chronic pancreatitis using natural language processing of radiology reports. PLoS ONE 15(8): e0236817. https://doi.org/10.1371/journal.pone.0236817

[3] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–1240.

[4] Devlin J, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT. 2019:4171-4186.

[5] Yao, L., Mao, C. & Luo, Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Med Inform Decis Mak 19 (Suppl 3), 71 (2019). https://doi.org/10.1186/s12911-019-0781-4