JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 6, pp. 11-20, June 2025
https://doi.org/10.9708/jksci.2025.30.06.011

Improving Deep Learning Performance on Imbalanced Medical Data Using Natural Language Data Augmentation Technique

Tae-Hyeong Kwon*, Dae-Ho Kim*, Se Young Kim**, Ok-Ran Jeong***

*Student, School of Computing, Gachon University, Seongnam, Korea
**Professor, Department of Nursing, Changwon National University, Changwon, Korea
***Professor, School of Computing, Gachon University, Seongnam, Korea

[Abstract]

In this study, we developed a model to support nurse decision-making using Korean nursing record data and explored methods to enhance performance by applying data augmentation techniques. Previous research primarily focused on English medical data, resulting in a lack of studies on Korean medical data. To address this gap, we utilized electronic medical record (EMR) data from abdominal surgery patients and developed a KoBERT-based model for predicting nursing actions. Additionally, we applied techniques such as up/down sampling, few-shot augmentation, back-translation, and synonym replacement to mitigate data imbalance and compared their performance. Experimental results show that the Few-shot Augmentation achieved the highest performance, confirming that data augmentation is effective in increasing the diversity of EMR data.

▸ **Key words**: Data Augmentation, Few-shot, Back-Translation, Synonym Replacement

[요 약]

본 연구에서는 한국어 간호기록 데이터를 활용하여 간호사의 의사결정을 지원하는 모델을 개발하고, 데이터 증강 기법을 적용하여 성능을 향상시키는 방안을 탐구하였다. 기존 연구들은 영어 의료 데이터를 중심으로 이루어져 한국어 의료 데이터에 대한 연구가 부족한 상황이었다. 이를 해결하기 위해 복부 수술 환자의 전자의무기록(EMR) 데이터를 활용하고, KoBERT 기반의 간호 조치 예측 모델을 개발하였다. 또한, 데이터 불균형 문제를 완화하기 위해 Up/down sampling, Few-shot Augmentation, Back-translation, Synonym Replacement 등의 기법을 적용하고 성능을 비교 분석하였다. 실험 결과, Few-shot Augmentation 기법이 가장 높은 성능을 기록하였으며, 이를 통해 데이터 증강이 EMR 데이터의 다양성을 높이는 데 효과적임을 확인하였다.

▸ **주제어**: 데이터 증강, 퓨샷 증강, 역번역, 동의어 대체

• First Author: Tae-Hyeong Kwon, Corresponding Author: Se Young Kim, Ok-Ran Jeong
*Tae-Hyeong Kwon (kth2642@gachon.ac.kr), School of Computing, Gachon University
*Dae-Ho Kim (ikimdh91@gachon.ac.kr), School of Computing, Gachon University
**Se Young Kim (sarakimk@changwon.ac.kr), Department of Nursing, Changwon National University
***Ok-Ran Jeong (orjeong@gachon.ac.kr), School of Computing, Gachon University
• Received: 2025. 04. 03, Revised: 2025. 05. 07, Accepted: 2025. 05. 28.

Copyright © 2025 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945
www.kci.go.kr

I. Introduction

의료 현장에서는 인공지능 기술의 발전과 함께 방대한 환자 데이터를 활용한 임상 의사결정 지원시스템(CDSS)에 대한 관심이 높아지고 있다. 특히 간호사는 한국 보건의료 인력의 약 19.5%를 차지하면서도 환자 상태 모니터링과 판단 업무에 중요한 역할을 담당하는 핵심 인력이다[1]. 고령 인구 증가와 만성질환자의 급증으로 환자 감시와 신속한 판단의 중요성이 더욱 커졌지만, 간호 인력의 부족과 기록, 인수인계와 같은 간접간호 업무 비중 증가로 인해 간호사의 업무 부담은 지속적으로 증가하고 있다. 이러한 배경 아래에서 디지털 헬스 기술과 AI를 간호 영역에 도입하여 간호사의 감시 및 판단 업무를 지원하고 업무 효율성을 높이기 위한 노력이 필요하다. AI 기반 CDSS는 간호사가 의사결정을 내릴 때 필요한 정보를 쉽게 빠르게 제공하여 간호 업무 부담을 경감하고, 환자 간호의 질을 향상할 수 있다[2].

그러나 현존하는 의료 AI 연구의 상당수가 영어권 임상 데이터에 집중되어 있어 한국어 의료 데이터의 활용에 한계가 있다. MIMIC-IV 등과 같은 대규모 공개 의료 코퍼스의 대부분이 영어로 구성되어 있고, 기존의 BERT와 같은 사전 학습 언어모델은 주로 영어 대규모 코퍼스를 기반으로 개발되었기 때문에 의료 도메인에 특화되거나 한국어의 형태론적 특성을 충분히 반영하지 못한다는 한계가 있다[3]. 이를 보완하기 위해 한국어 형태소와 어휘를 반영하여 사전 학습된 KoBERT와 같은 한국어 특화 언어모델이 개발되었으나, 여전히 의료 도메인 특화 학습이 부족하거나 한국어 의료 문서의 구조적 특성을 반영하는 연구는 상대적으로 드문 실정이다[3]. 또한 국내 전자의무기록(EMR)의 간호기록 데이터는 한글과 영문이 혼영 되어 사용되는 경우가 많아 일반적인 자연어 처리 도구로 처리하기에 어려움이 따른다[4]. 한국어 형태소 분석기는 의료 분야 용어를 충분히 인식하지 못하기 때문에 한영 혼용 텍스트를 전처리하는 과정에서 정보 손실이 발생하는 문제도 보고되었다[3]. 이러한 언어 장벽과 데이터 부족 문제로 인해 한국어 간호 기록을 활용한 AI 연구는 아직 현저히 부족하며[1] 영어 중심으로 축적된 지식을 그대로 적용하기 어려운 것이 현실이다.

본 연구는 이러한 한계를 극복하고자 실제 의료기관에서 수집된 한국어 간호 기록 데이터를 활용하여 CDSS 예측 모델을 구축하는 것을 목표로 한다. 국내 간호기록에서 널리 사용되는 DAR(Data, Action, Response) 형식을 바탕으로, 환자 상태(D)에 따른 간호활동(A)을 예측하는 작업을 설계하였다. 이를 위해 한국어 텍스트 처리에 적합한 사전 학습 모델인 KoBERT를 기반으로 모델을 설계하고, 데이터 증강 기법(Data Augmentation)을 적용하여 데이터의 다양성과 학습 안정성을 확보하였다. 다만 실제 간호 현장에서는 하나의 환자에 대해 복수의 간호 활동이 연속적으로 시행되는 경우가 많지만, 본 연구에서는 학습 가능한 구조를 설계하기 위해 단일 중재 예측을 중심으로 모델을 구성하였다.

II. Related Works

1. Nursing Surveillance

간호중재분류체계(NIC)에서 간호감시는 “임상 의사결정을 위한 환자 데이터의 지속적이고 목적적인 수집, 해석 및 통합”으로 정의된다[5]. 이러한 감시 활동은 간호사가 환자의 상태 변화를 초기에 포착하고 위험을 판단하는 핵심 역량으로 간주된다. 실제로 간호사는 임상 현장에서 환자와 가장 오랜 시간 밀접하게 상호작용하며, 합병증 예방과 이상 징후 탐지에 있어 최적의 위치에 있다는 보고가 있다[6]. 특히 숙련된 간호사는 활력징후, 검사 결과, 임상 징후 등 다양한 정보를 종합하여 데이터의 패턴과 의미를 해석함으로써 환자의 위험 수준을 분류하고 선제적으로 대응한다.

그러나 임상에서 간호사에 의한 수작업 감시에는 한계가 존재한다. 환자 감시 업무는 인력 부족과 업무 부담으로 인해 지속성과 정확성에 영향을 받을 수 있으며, 주관적 판단에 의존하기 쉽다. 이러한 한계를 보완하기 위해 초기 경보 점수(Early Warning System, EWS)가 도입되어 활력징후 기반의 점수를 통해 악화 가능성을 객관적으로 경고하고 있다[7]. 더 나아가 스페인의 Catalonia 병원에서는 간호사 주도의 환자 모니터링 체계인 VIDA 시스템을 개발하여 환자 위험도를 5단계로 분류하고, 조기 악화 징후를 체계적으로 감시하고 있다[8]. EWS, VIDA와 같은 시스템은 간호사의 위험 판단을 지원하는 실제 사례를 보여주며, 간호 감시의 실무적 한계를 보완하는 기술적 도구의 필요성을 보여준다.

2. EMR(Electronic Medical Record)

의료 현장의 기록 체계는 종이 문서에서 전자의무기록(EMR)으로 급속히 전환되며 환자 정보의 디지털화가 이루어졌다. 간호기록 역시 다양한 임상 관찰과 중재 내용이 EMR에 구조화 또는 비구조화된 형태로 저장되고 있다. 하

www.kci.go.kr

지만 한국어 임상 문서는 언어적 특성상 분석이 복잡하여 전처리에 어려움이 따른다. 한국어 의료 데이터는 교착어적 특성과 의료 분야의 복잡한 전문용어로 인해 기계 처리 난이도가 높으며[3], 간호기록에는 한글 약어, 영어 혼용, 축약 표현 등이 빈번하여 자연어 처리 과정에서 특수한 난제를 제시한다. EMR 데이터의 축적은 임상 의사결정 지원을 위한 인공지능 연구를 활성화하였다. 실제로 대규모 환자 기록을 활용해 임상 결과를 예측하거나 의료진에게 조언을 제공하는 다양한 연구가 국내외에서 보고되었다[9][10].

3. CDSS with AI

임상 의사결정 지원시스템(CDSS)은 환자 진료의 예방, 진단, 처치, 투약, 예후 등의 단계에서 임상의의 의사결정을 보조하도록 설계된 지식 기반 상호작용 시스템을 말한다[11]. CDSS는 환자로부터 얻어진 임상 정보를 분석하여 적시에 경고나 권고안을 제시함으로써 의료인의 판단을 지원한다. 예를 들어 약물 처방 시 부작용 또는 상호작용을 자동으로 확인해 경고하거나, 환자 상태에 따른 진단 가능성을 제안하는 등 다양한 형태의 CDSS가 개발되어 임상 현장에 도입되어 왔다. 이러한 시스템은 의료 지식베이스와 환자 데이터의 연계를 통해 의료진의 인지 부하를 줄이고 진료의 안전성과 효율성을 높이는 것을 목적으로 한다. 그러나 지금까지의 CDSS 개발은 주로 의사를 대상으로 한 진단, 처방 지원에 초점을 맞추고 있어 간호사의 임상 판단을 직접 보조하는 사례는 상대적으로 드물다. CDSS 활용이 환자 안전과 간호 업무 효율을 향상할 수 있음에도 불구하고 간호 분야에서의 도입은 제한적이며, 실제 구현의 성공률도 낮은 것으로 보고된다[12].

4. Medical Data Augmentation

의료 도메인 자연어처리(NLP) 과제에서는 데이터 부족으로 인한 모델 학습 한계가 빈번한 문제로 대두된다. 이를 보완하기 위해 다양한 데이터 증강(Data Augmentation) 기법이 연구되고 있다. 데이터 증강은 기존 데이터로부터 새로운 학습 샘플을 생성해 데이터의 양과 다양성을 늘리는 방법으로, 최근 임상 텍스트 분류나 개체 인식 등 의료 NLP 작업의 성능 향상을 위해 적극 활용되는 추세이다. 대표적인 텍스트 증강 기법으로 역번역(Back-Translation)과 동의어 치환(Synonym Replacement)이 있다.

역번역은 원문 문장을 다른 언어로 번역한 뒤 다시 원래 언어로 역번역하여, 의미는 유지하되 표현이 다른 문장을 생성하는 방법이다. 이 기법은 문장 단위로 데이터 다양성을 크게 높여주지만, 단순 역번역 결과는 문법 오류 등 품질 저하 문제가 있어 무분별한 사용 시 모델 성능에 악영향을 줄 수 있다. 이러한 한계를 보완하기 위해 최근에는 역번역으로 생성된 문장을 판별기로 거르거나, 다국어 번역을 활용하는 등 품질을 향상시키는 변형 기법들도 제안되고 있다[13]. 최근 연구에서는 의료 도메인의 Aspect Term Extraction(ATE) 작업을 개선하기 위해 역번역 기법을 적용하였다[14]. 해당 연구에서는 프랑스어, 중국어, 독일어를 중간 언어로 활용하여 의료 텍스트를 역번역한 후, 번역된 텍스트를 원문과 비교하여 일관성을 유지하는 방법을 도입하였다. 역번역 기법을 적용한 데이터셋을 학습한 모델은 ATE 성능이 최대 3.46% 향상되며 역번역 증강 기법을 통해 다양한 문맥에서의 표현을 보다 효과적으로 학습할 수 있음을 보였다.

동의어 치환은 문장 내 특정 단어를 동일한 의미의 다른 표현으로 바꾸어 새로운 문장을 만드는 방식으로, 의료 용어 사전이나 UMLS 같은 지식 그래프를 활용해 의미를 보존하면서 표현을 다양화한다. 의료 도메인에서는 Named Entity Recognition(NER) 성능을 개선하기 위해 동의어 치환 기법을 적용한 연구 사례가 있다[15]. 연구진은 Easy Data Augmentation(EDA) 기법을 활용하여 의료 데이터셋에서 특정 의료 용어를 해당 도메인에서 자주 사용되는 동의어로 치환하는 방법을 제안하였다. 실험 결과, 동의어 치환 기법을 적용한 NER 모델이 기존 모델보다 높은 성능을 기록하며, 데이터 다양성을 확보하는 데 효과적인 방법임을 입증하였다.

또 다른 데이터 증강 방법으로는 Few-shot Augmentation 기법이 있다. Few-shot 기반 데이터 증강은 소량의 예시만으로 사전 학습 언어모델을 조정하거나 프롬프트에 삽입해 라벨이 붙은 합성 문장을 대량 생성함으로써 학습 데이터의 양과 다양성을 동시에 확보할 수 있는 증강 기법이다[16][17]. 최근 ChatGPT로 생성한 Few-shot 증강 문장과 다중 스케일 추출을 결합해 Few-shot NER 성능을 높인 연구가 발표되었다[18]. 해당 연구에서는 PubMedBERT 기반 실험에서 BC5CDR-Disease 등 4개의 의료 코퍼스의 5-shot 설정에서 F1-Score를 최대 10.2%까지 향상시켰다.

www.kci.go.kr

III. Methodology

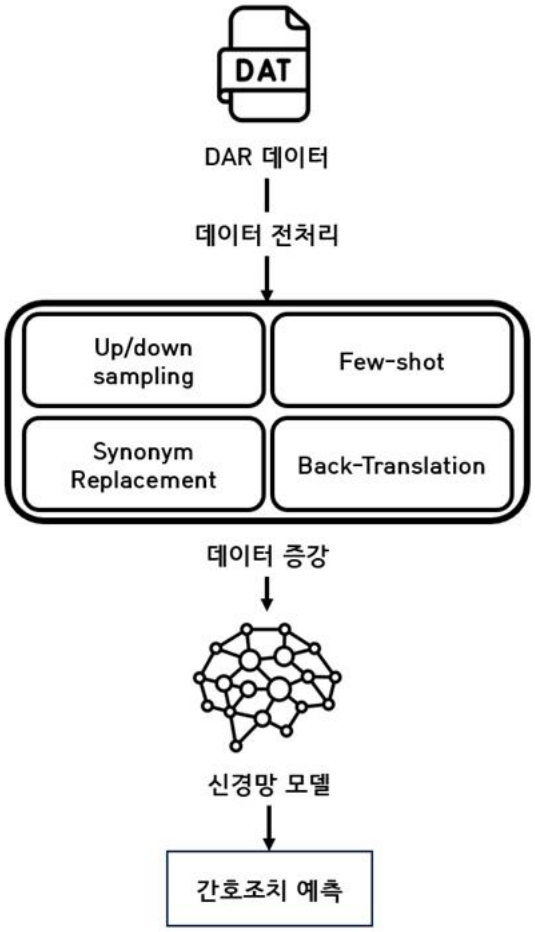

Fig. 1. CDSS Model Development Pipeline

[Fig. 1]은 본 연구의 CDSS 개발 파이프라인을 나타낸 것이다. 먼저 DAR 간호기록에 대한 데이터 전처리를 진행하고, 증강 기법을 적용한 데이터를 통해 신경망 모델을 학습시켜 간호 활동을 예측하고 임상 의사결정을 지원하도록 한다.

1. Data Preprocessing

불필요한 공백, 줄 바꿈, 특수문자 등으로 인해 A 클래스 수가 불필요하게 늘어나는 문제를 줄이고자 정규표현식을 사용해 모든 기록 문장의 양 끝 공백이나 줄 바꿈, 문자 등을 제거하였다. 하나의 D에는 여러 개의 A가 따라올 수 있으며, 전처리 후 평균적으로 따라오는 A의 개수는 1.23개이다. [Table 1]은 DAR 간호기록의 예시로, 동일한 D에 대해 단일 조치로 기록된 경우와, 다중 조치로 기록된 경우가 모두 존재하는 것을 확인할 수 있다. 하지만 현재 데이터에서 활용할 수 있는 변수는 DAR 항목 하나뿐이기 때문에 현실적으로 다중, 단일 조치를 구분하여 학습하기 어렵다는 문제가 존재했고, 이를 위해 각 D에 가장 자주 나타난 A를 해당 D에 대한 대표 조치로 지정하고, 빈도가 같을 경우 둘 중 먼저 기록된 A를 선택하여 예측 문제를 단일 조치 형태로 단순화하였다. 최종적으로 데이터는 환자가 어떤 증상을 보이는지 서술된 D 데이터 하나와 해당 증상에 대해 간호사가 어떤 조치를 했는지가 서술된 A 데이터 하나의 쌍으로 구성된다. [Table 2] 는 최종 데이터셋 쌍의 예시를 보여준다.

Table 1. Nursing Record DAR Data Example

DAR	Contents
D	진단명:(Cholecystitis, unspecified),주증상:(복통으)로 (선제격리)위해 입원함
A	간호요구에 대한 간호계획을 본인과 보호자에게 설명함
A	담당 간호사를 소개하고 병실 안내함. 환자, 보호자에게 입원생활 안내문 제공하고, 입원생활(병실구조 및 편의시설,환자권리보호,불편신고체계, 비상시 대피요령 등)에 대하여 설명함
A	간호사 호출기와 침상난간 사용법,낙상주의,도난사고 예방,화재,비상시 대피방법에 대해 교육함
A	보호자 상주에 대해 설명함
D	진단명:(Cholecystitis, unspecified),주증상:(복통으)로 (선제격리)위해 입원함
A	간호요구에 대한 간호계획을 본인과 보호자에게 설명함

Table 2. Nursing Record DAR Data Example After Preprocessing

D	A
진단명:(Cholecystitis, unspecified),주증상:(복통으)로 (선제격리)위해 입원함	간호요구에 대한 간호계획을 본인과 보호자에게 설명함

2. Data Augmentation

각 A의 클래스들의 등장 빈도에 대한 평균은 15.17, 표준편차는 768.07로 굉장히 극단적인 데이터 불균형을 갖고 있다. 특히 일부 클래스는 데이터 샘플이 매우 많거나, 매우 적어 모델의 학습 및 일반화 능력에 부정적인 영향을 미칠 수 있었다. 데이터셋에서 A 데이터 빈도수 상위 5개 클래스의 비율을 조사한 결과는 [Table 3]과 같다. 결과에서 나타난 것과 같이 상위 5개의 클래스가 전체 데이터셋의 52.9%를 차지할 정도로 A 클래스 데이터의 불균형이 극단적으로 심한 상태였다.

www.kci.go.kr

Table 3. Top 5 frequency A data classes

Contents	Percentage of total dataset
"통증을 표현하도록 격려함"	21.34%
"통증 증가시 간호사에게 알리도록 교육함"	16.47%
"통증이 심해지기 전에 필요시 진통제를 요구하도록 교육함"	11.05%
"간호 요구에 대한 간호 계획을 본인에게 설명함"	2.58%
"담당 간호사를 소개하고 병실 안내함 . 환자 , 보호자에게 입원 생활 안내문 제공하고 , 입원 생활 병실 구조 및 편의 시설 , 환자 권리 보호 , 불편 신고 체계 , 비상 시 대피 요령 등 에 대하여 설명함"	1.46%

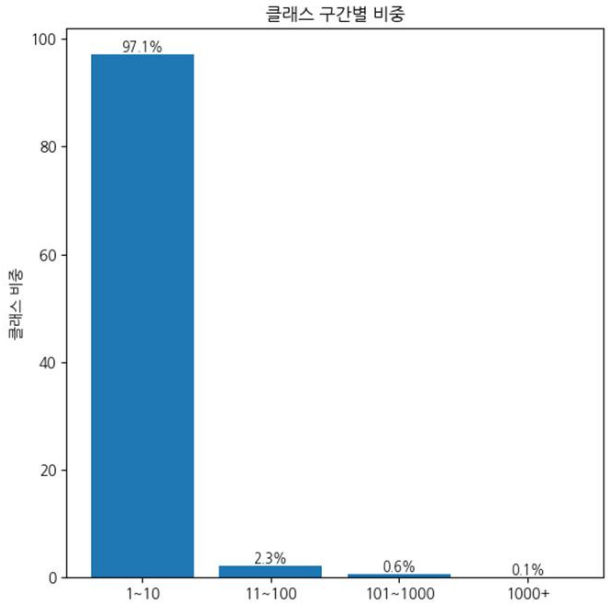

Fig. 2. Weight by A class count range

이러한 데이터 불균형 문제를 해소하고 모델의 학습 효율성을 높이기 위해, 샘플 수가 적은 희소 A 클래스에 대해, 해당 클래스와 매핑된 D 문장들을 증강 대상으로 선정하여 증강 기법을 적용하고 D-A 쌍을 추가 생성하였다. [Fig. 2]는 A 클래스 발생 빈도를 로그 스케일로 나누어 각 구간에 포함된 샘플 비중을 그래프로 나타낸 것으로, 데이터가 10개 이하인 클래스가 전체 클래스 중 97.1%를 차지할 만큼 A 클래스 대부분은 소수 클래스로 구성되어 있음을 보여준다. 이러한 통계적 근거를 바탕으로 본 연구에서는 10개 이하의 샘플 데이터를 보유한 희소 A 클래스와 쌍을 이루는 D 문장들에 데이터 증강을 적용하였다.

본 연구에서는 데이터 불균형 완화 및 모델 성능 향상을 위해 Up/down Sampling, Few-shot Augmentation, Back-Translation, Synonym Replacement의 네 가지 데이터 증강 기법을 적용하였다. [Table 4]에서는 각 기법의 정의, 적용 기준 및 적용 방식을 확인할 수 있다.

Up/down Sampling은 특정 A 클래스의 샘플 수가 다른 클래스에 비해 현저히 적거나 많을 경우, 과소표본추출(under-sampling) 및 과대표본추출(over-sampling)을 수행하여 데이터의 균형을 맞추었다. 소수 클래스의 경우 샘플 개수가 설정한 최소 임계값(10개)보다 적을 경우 해당 클래스를 가진 D-A 샘플 데이터 쌍을 랜덤 복제하여 부족한 샘플을 추가 생성하였다. 반대로 다수 클래스의 경우 샘플 개수가 설정한 최대 임계값(100개)을 초과하는 경우 해당 클래스에서 무작위로 일부 샘플을 추출하여 샘플 수를 제한하였다.

Few-shot Augmentation은 Google의 사전 학습된 다국어 모델인 mT5(Multilingual T5)[19]를 활용하여 수행되었다. 본 연구에서는 특정 D-A 쌍을 입력으로 주어 새로운 D 문장을 생성하도록 하였다. 프롬프트를 통해 모델에게 기존 D 문장과 A 문장을 제공하고, A 문장에 적절한 새로운 D 문장을 생성하여 새로운 D-A 쌍을 생성하고, 이를 원본 데이터와 결합하여 최종 학습 데이터셋을 구성하였다.

Back-Translation은 원본 D 문장을 영어로 번역한 후 다시 한국어로 번역하는 과정을 통해 이루어졌으며, 한국어에서 영어로 번역하는 과정에서 Helsinki-NLP의 Marian MT 모델[20]을, 영어에서 한국어로 역번역하는 과정에서 Facebook의 M2M100 모델[21]을 사용하였다.

Synonmy Replacement는 Few-shot Augmentation과 동일하게 사전학습된 mT5 모델을 통해 수행하였다. 각 D 문장에서 랜덤하게 선택된 20%의 단어를 마스크 처리한 후, mT5 모델이 이를 대체할 단어를 생성하도록 하였다. mT5 모델은 마스크 처리된 문장을 입력받아 대체할 단어 후보를 생성하고, 상위 5개 후보 중 하나를 무작위로 선택하여 원래 문장의 단어를 치환했다.

3. CDSS with KoBERT

본 연구는 한국어 의료 도메인 간호 기록에 데이터 증강 기법을 적용하고, KoBERT 기반 모델에 학습시켜 간호사의 임상 의사결정을 예측하도록 하였다. KoBERT는 구글의 BERT[22] 모델을 기반으로 한국어 대규모 코퍼스에 사전 학습된 모델이다. 모델은 환자의 증상 데이터를 입력받아 가장 적절한 간호 조치를 예측하는 방식으로 설계되었

www.kci.go.kr

Table 4. Augmentation Methodology Summary

Augmentation	Definition	Application Criteria	Application Method
Up/down Sampling	데이터의 균형을 맞추기 위해 과소표본추출 및 과대표본추출 수행	샘플 개수가 최소 임계값 미만이거나 최대 임계값 초과 시	소수 클래스 랜덤 복제, 다수 클래스 랜덤 추출
Few-shot Augmentation	소량의 예시로 사전학습 언어모델을 조정하여 새로운 학습 샘플 생성	샘플 개수가 10개 이하인 희소 A 클래스	mT5 모델에 D-A 쌍을 입력으로 주어 새로운 D 문장 생성
Back-Translation	문장을 다른 언어로 번역 후 다시 원래 언어로 변환하여 데이터 다양성 확보	샘플 개수가 10개 이하인 희소 A 클래스	한국어에서 영어로 번역 (MarianMT) 후 영어에서 한국어로 역번역(M2M100)
Synonym Replacement	기존 텍스트에서 특정 단어를 의미가 유사한 동의어로 대체	샘플 개수가 10개 이하인 희소 A 클래스	임의 단어 마스킹 후 mT5 모델로 동의어 생성 및 치환

다. 이를 위해 KoBERT에 D 데이터를 입력으로 주어 임베딩 벡터로 변환한 후, 다중 신경망을 통해 최종적으로 A 클래스를 예측하는 모델을 구축하였다. [Fig. 3]은 KoBERT Classifier의 구조를 나타낸 그림이다. KoBERT를 통해 768차원의 임베딩 벡터로 변환된 문장을 Linear 레이어를 거쳐 512차원의 표현 공간으로 변환한 후, 비선형 변환을 위해 ReLU 함수를 적용하였다. 이후 과적합을 방지하기 위해 Dropout 레이어를 추가하였고, Dropout 값은 0.3으로 설정하였다. 이후 최종적으로 Linear 레이어를 통해 A 클래스를 예측한다. 모델의 가중치 최적화에는 Adam Optimizer를 사용하였으며, 손실 함수는 Cross-Entropy Loss를 사용했다.

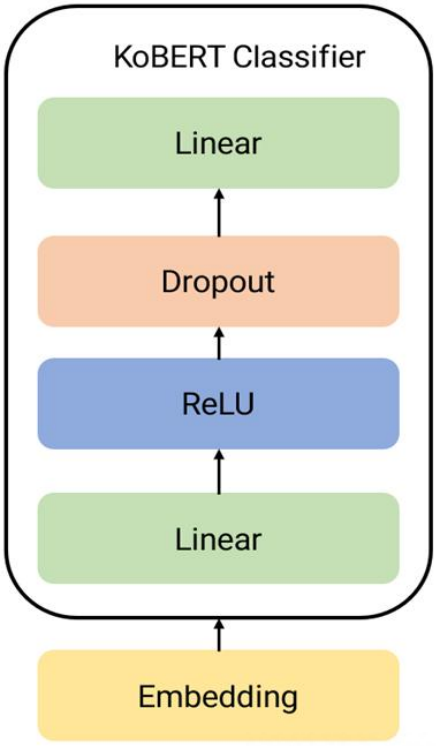

Fig. 3. KoBERT Classifier

IV. Experiment and Result

1. Dataset

본 연구에서 사용된 데이터는 경남 지역 A 상급종합병원에서의 EMR 데이터와 전자 간호기록 데이터로, 2018년 10월 1일부터 2023년 9월 30일까지 일반외과 병동에 입원하여 복부 수술을 받은 약 8,500명의 환자를 대상으로 수집한 데이터들이 포함되어 있다. 해당 EMR 데이터셋은 S 대학병원 IRB 승인과 데이터심의위원회의 승인을 받은 뒤, 병원정보시스템 및 EMR 담당 부서와 협의를 거친 후 사용되었다. EMR 데이터셋은 총 16개의 데이터 테이블로 구성되어 있고, 본 연구에서는 그 중 DAR 간호기록 테이블만을 사용하였다. DAR 간호기록 데이터는 환자의 식별번호, 기록 일자, 기록시간, 환자의 증상 호소 내용, 상태 등이 기록된 D(Data), 그에 대한 간호사의 활동이 기록된 A(Action), 활동 이후 환자의 징후를 기록한 R(Result) 데이터로 구성되어 있다. 모든 기록 내용은 텍스트 데이터로 구성되어 있다. D 데이터와 R 데이터에는 환자가 느끼는 고통의 정도를 표현한 NRS 수치가 함께 기록되어 있다. 하지만 이는 기구를 통해 객관적인 값을 측정한 것이 아닌, 환자의 주관적인 기준에 의해 기록된 수치이며 대부분의 케이스에서 간호사의 조치 이후 NRS 수치의 변동이 없었기 때문에 본 연구에서는 D 데이터와 A 데이터만을 활용하였다. 최종적으로 사용한 DAR 간호기록 데이터셋은 총 230,406개의 D-A 쌍으로 구성되어 있으며, 학습, 검증, 테스트 데이터셋은 각각 161,284개, 34,561개, 34,561개의 데이터로 구성되었다.

2. Experiment Environment & Metric

본 연구의 실험 환경은 GeForce RTX 2080 SUPER 16GB를 통해 수행되었다. Batch size는 32, 학습률은

www.kci.go.kr

2e-5, epoch 수는 10으로 설정하였으며, 모델의 성능 평가 지표로 Accuracy, Precision, Recall, F1-Score를 사용하였다. 본 연구에서는 데이터 전처리 단계에서 각 D에 대해 가장 빈도수가 높은 A를 대표 조치로 설정하였다, 따라서 모델의 정확도는 입력된 D에 대해 미리 정의된 단일 정답 A를 얼마나 정확하게 예측하는지를 기준으로 평가된다.

3. Single Augmentation Result

Table 5. Single Augmentation Result

Model	Acc	Prec	Recall	F1
w/o Augmentation	0.7360	0.6949	0.7360	0.7003
Up/down sampling	0.7997	0.7768	0.7997	0.7746
Few shot	**0.9028**	**0.8898**	**0.9028**	**0.8867**
Back translation	0.8348	0.8464	0.8348	0.8208
Synonym replacement	0.7765	0.7839	0.7765	0.7611

[Table 5]는 단일 증강 기법 적용 시 실험 결과를 나타낸 표로, w/o Augmentation은 데이터 증강을 적용하지 않고 예측을 수행했을 때를 의미한다. 실험 결과, Few-shot Augmentation 기법이 Accuracy 90.28%, Precision 88.98%, Recall 90.28%, F1-Score 88.67%를 기록하며 가장 우수한 성능을 보였다. 이는 Few-shot Augmentation이 희소 클래스에 대해 맥락적으로 적절한 문장을 생성함으로써 모델이 데이터가 부족한 특정 간호 조치 유형에 대해서도 견고하게 학습할 수 있는 기반을 마련했음을 의미한다. Back-Translation 기법 또한 Accuracy 83.48%를 기록하며 데이터 표현의 다양성을 높이는 데 기여하였다. 원문의 의미를 유지하면서도 다양한 표현을 학습할 수 있도록 지원하는 특성이 모델 성능 향상이 긍정적인 영향을 미친 것으로 해석할 수 있다. 이는 모델이 동일한 의료 개념을 다양한 언어적 형태로 인식하고 처리할 수 있도록 하는 데 기여했음을 의미한다. 반면, Up/down sampling 기법은 Accuracy 79.97%를 기록하며 불균형 데이터를 보완하는 데 유의미한 효과를 보였지만, 극단적인 클래스 불균형을 해결하고 데이터의 질적 다양성을 확보하는 데는 한계가 있었다. Synonym Replacement 기법은 77.65%의 Accuracy로, 상대적으로 낮은 성능을 기록하였다. 이는 mT5 모델과 같은 일반적인 언어 모델이 한국어 의료 도메인 지식을 충분히 포함하지 못하는 한계점으로 기인한 것으로 분석된다. 일반적인 한국어 코퍼스로 학습된 모델은 “통증”과 “고통”과 같은 일반적인 동의어는 잘 치환할 수 있지만, “복부 수술”과 같은 특정 의료 용어는 그 문맥적 의미를 정확히 이해하고 적절한 동의어를 찾아내기 힘들다. 또한 본 연구에서 사용한 데이터의 경우 “우리한”과 같은 지방 방언 표현들도 포함되어 있었기 때문에, 한국어에 특화된 사전 학습 모델이 아닐 경우 이러한 지역의 언어적 특성을 반영해야 하는 용어의 치환은 의미 왜곡을 초래하거나, 실제 임상에서 사용되지 않는 부적절한 표현을 생성하여 데이터에 노이즈를 추가할 가능성이 있다. 이러한 현상은 결과적으로 모델이 의미론적으로 부정확하거나 문맥에 맞지 않는 증강된 데이터를 학습하게 되어 성능 저하로 이어질 수 있다.

4. Combination Augmentation Result

데이터 증강 기법들의 조합이 모델 성능에 미치는 영향을 추가로 분석하기 위해, 여러 증강 기법을 함께 적용한 추가 실험을 진행하였다. 조합 실험 결과는 [Table 6]과 같다. 실험은 첫 번째 증강 기법을 적용한 데이터에 이어지는 데이터 증강 기법을 적용하는 방식으로 수행하였다. 예를 들어, Few-shot + Back Translation의 경우, Few-shot Augmentation을 적용하여 생성한 D-A 쌍 데이터의 D 데이터에 다시 Back-Translation을 적용하여 문장을 한 차례 더 변형하는 식으로 진행했다. 실험 결과, 모든 조합 기법이 단일 Few-shot Augmentation 기법보다 낮은 성능을 기록하였다. 이는 Few-shot Augmentation이 단독으로도 매우 강력한 증강 효과를 제공하며, 다른 기법과의 조합이 반드시 추가적인 성능 향상으로 이어지지 않을 수 있음을 의미한다.

Table 6. Combination Augmentation Result

Model	Acc	Prec	Recall	F1
Few shot + Back Translation	0.8632	0.8550	0.8617	0.8583
Few shot + Synonym Replacement	0.8352	0.8290	0.8370	0.8330
Back Translation + Synonym Replacement	0.7924	0.7874	0.7819	0.7846

5. Qualitative Analysis and Error Cases

본 연구에서 사용한 EMR 데이터의 경우 경남 지방 방언이 함께 포함된 간호 기록이 종종 존재한다. 예를 들어

www.kci.go.kr

“우리한”이라는 표현은 “신체의 일부가 몹시 아리고 욱신욱신한 느낌이 있다”는 뜻의 경상 지방 방언으로, 해당 표현이 사용된 간호 기록이 상당수 존재하는 것을 확인할 수 있었다. Few-shot Augmentation의 경우 “우리한”이라는 표현의 의미를 모르더라도 주어진 D와 A의 맥락을 파악한 후 해당 표현의 의미를 추론하고 유사한 의미를 가진 D를 비교적 안정적으로 생성하는 경우가 많았지만, 주어진 문장을 번역하는 Back-Translation의 경우 해당 표현을 제대로 번역하지 못하는 경우가 많았고, Synonym Replacement의 경우 “우리한”을 “유리한”, “우려한” 등으로 치환하는 등 문장의 의미가 완전히 왜곡된 형태로 증강 데이터가 생성되는 오류가 빈번하게 발생하였다.

6. Imbalance Mitigation Analysis

데이터 불균형 문제를 보완하기 위한 증강 기법의 효과를 더욱 정밀하게 평가하기 위해 micro F1-Score 외에도 macro F1-Score를 추가로 분석하였다. [Table 7]에 따르면 증강 없이 학습한 모델의 경우 macro F1은 0.6092로 micro F1 대비 약 0.09 낮은 값을 기록하였으며, 이는 다수 클래스 예측 성능이 집중된 불균형한 분포를 반영한다. 모든 증강 기법은 micro와 macro F1에서 모두 성능 향상을 보였으며, 특히 Few-shot 증강은 두 지표 모두에서 가장 높은 성능을 기록했다. 그뿐만 아니라 micro F1과 macro F1 간의 차이가 약 0.05로, 모든 증강 기법 중에서도 가장 적은 차이를 보였다. 이는 단순히 전체 예측 정확도를 높이는 데 그치지 않고, 소수 클래스에 대한 예측 성능까지 효과적으로 개선했음을 보여준다.

Table 7. micro, macro F1-Score by model

Model	micro F1	macro F1
w/o Augmentation	0.7003	0.6092
Up/down Sampling	0.7746	0.6984
Few-shot	**0.8867**	**0.8343**
Back Translation	0.8208	0.7510
Synonym Replacement	0.7611	0.6644

V. Discussion

본 연구는 한국어 간호 기록 데이터에 데이터 증강을 적용하여 간호사의 임상적 의사결정을 지원하는 모델의 성능을 향상하는 데 기여했으나 여전히 개선해야 할 여러 한계점들이 남아있다. 실험은 하나의 D 데이터에 대해 하나의 A 데이터만을 매핑하는 단일 활동 예측 방식으로 진행되었는데, 이는 실제 간호 현장에서 하나의 증상에 대해 복수의 간호 활동이 연속적으로, 또는 복합적으로 시행되는 실무와의 괴리가 있다. 현재 데이터에서 활용할 수 있는 변수가 DAR 항목에 국한되어 있어 다중 조치 상황을 구분하기 어려운 문제로 인해 가장 빈도수가 높은 A를 정답으로 정의했지만, 해당 방식에 대한 의학적 타당성이 불충분하다는 문제도 있다. 이러한 한계점을 극복하기 위해 향후 연구에서는 DAR 간호기록과 다른 테이블을 함께 사용하여 하나의 증상에 대해 여러 간호활동을 동시에 예측하는 multi-label classification이나 순차적으로 필요한 활동을 예측하는 sequence-to-sequence 방식의 간호감시 모델 개발을 고려할 수 있다. 또한 최근 한국어를 지원하면서도 의료 도메인에 특화되어 파인 튜닝된 Med-Gemini[23]와 같은 LLM의 적용을 고려할 수 있다.

VI. Conclusion

본 연구에서는 한국어 간호 기록 데이터에 데이터 증강 기법을 적용하여 해당 기법이 임상 의사결정 지원 모델(CDSS)의 개발에 있어 어떤 기여를 줄 수 있는지 연구하였다. 의료 데이터의 희소성과 클래스 불균형이라는 문제를 해결하기 위해 Up/down sampling, Few-shot Augmentation, Back-Translation, Synonym Replacement 기법을 적용하고 KoBERT를 통해 그 효과를 비교 분석하였다. 실험 결과, Few-shot Augmentation 기법이 모든 평가 지표에서 가장 높은 예측 수치를 기록하며, 한국어 의료 도메인에서의 데이터 증강의 중요성을 입증하였다. 이러한 결과는 한국어 의료 환경에서 데이터 부족 문제를 극복하고 임상 의사결정 지원 모델의 예측 성능을 향상하는 데 데이터 증강이 필수적인 전략임을 의미한다.

본 연구를 통해 간호사의 의사결정 과정을 지원하고, 궁극적으로 의료 서비스의 효율성과 품질을 향상할 수 있는 기반을 마련하였다. 다만, 다중 조치가 빈번하게 필요한 실무 간호사 환경과 다르게 본 연구의 실험은 단일 조치 예측을 수행했다는 한계점이 있고, 이러한 한계점을 보완하기 위해 향후 연구에서 multi-label classification, sequence-to-sequence 모델, 그리고 LLM의 적용 등을 통해 지속적으로 개선해 나갈 수 있을 것으로 기대된다.

www.kci.go.kr

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00273954).

REFERENCES

[1] H.B. Lee, W.J. Moon, S.A. Kim, J.H. Lee, and O.J. Jang, "A Study on Exploring the Possibilities of AI Application for Improving Domestic Nursing Work," Journal of Korean Academy of Nursing Administration, Vol. 29, No. 5, pp. 564-576, December 2023. DOI: 10.11111/jkana.2023.29.5.564

[2] C. C. Halverson, and D. S. Tilley, "Nursing surveillance: A concept analysis," Nursing Forum, Vol. 57, No. 3, pp. 454-460, May 2022. DOI: 10.1111/nuf.12702

[3] Y. Kim, J.-H. Kim, J. M. Lee, M. J. Jang, Y. J. Yum, S. Kim, … and S. Song, "A pre-trained BERT for Korean medical natural language processing," Scientific Reports, Vol. 12, No. 13847, August 2022. DOI: 10.1038/s41598-022-17806-8

[4] C.H. Kim, "An Implementation of Natural Language Processing and Text Mining in Stroke Research," Journal of the Korean Neurological Association, Vol. 39, No. 3, pp. 121-128, July 2021. DOI: 10.17340/jkna.2021.3.2.

[5] G. M. Bulechek, H. K. Butcher, and J. M. Dochterman, Nursing Interventions Classification (NIC), 5th ed., Mosby, 2008.

[6] S. Dresser, "The role of nursing surveillance in keeping patients safe," Journal of Nursing Administration, Vol. 42, No. 7-8, pp. 361–368, July-August 2012.

[7] C. P. Subbe, M. Kruger, P. Rutherford, and L. Gemmel, "Validation of a modified early warning score in medical admissions," Quarterly Journal of Medicine, Vol. 94, No. 10, pp. 521–526, October 2001.

[8] J. Adamuz, M. González-Samartino, E. Jiménez-Martínez, et al., "Risk of acute deterioration and care complexity individual factors associated with health outcomes in hospitalised patients with COVID-19: a multicentre cohort study," BMJ Open, Vol. 11, No. e041726, February 2021. DOI: 10.1136/bmjopen-2020-041726

[9] A. Rajkomar, E. Oren, K. Chen, et al., "Scalable and accurate deep learning with electronic health records," NPJ Digital Medicine, Vol. 1, No. 18, May 2018. DOI: 10.1038/s41746-018-0018-y

[10] J.-M. Kwon, Y. Lee, Y. Lee, S. Lee, and J. Park, "An algorithm based on deep learning for predicting in-hospital cardiac arrest," Journal of the American Heart Association, Vol. 7, No. e008678, July 2018. DOI: 10.1161/JAHA.118.008678

[11] D.H. Lee, H.Y. Jung, M.H. Kim, M.E. Lim, D.H. Kim, Y.W. Han, Y.W. Kim, J.H. Choi, and S.H. Kim, "Clinical Decision Support System (CDSS) Technology Trends," Electronics and Telecommunications Trends, Vol. 31, No. 4, pp. 77-85, 2016.

[12] S. C. Lu, R. Brown, and M. Michalowski, "A clinical decision support system design framework for nursing practice," ACI Open, Vol. 5, No. 2, pp. e84–e93, April 2021. DOI: 10.1055/s-0041-1725902

[13] B. Shi, et al., "MDA: An intelligent medical data augmentation scheme based on medical knowledge graph for Chinese medical tasks," Applied Sciences, Vol. 12, No. 20, pp. 10655, October 2022. DOI: 10.3390/app122010655

[14] Q. Xu, Y. Hong, J. Chen, J. Yao, and G. Zhou, "Data augmentation via back-translation for aspect term extraction," Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, pp. 1–8, June 2023. DOI: 10.1109/IJCNN54540.2023.10191183

[15] A. M. Issifu, and M. C. Ganiz, "A simple data augmentation method to improve the performance of named entity recognition models in medical domain," Proceedings of the 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, pp. 763–768, September 2021. DOI: 10.1109/UBMK52708.2021.9558986

[16] Y. Meng, M. Michalski, J. Huang, Y. Zhang, T. Abdelzaher, and J. Han, "Tuning language models as training data generators for augmentation-enhanced few-shot learning," Proceedings of the 40th International Conference on Machine Learning, Vol. 202, pp. 24457–24477, July 2023.

[17] S. H. Lin, J. G. Zhu, C. J. Qian, J. J. Lin, J. Y. Deng, X. G. Wang, and T. Huang, "Data Augmentation using LLMs: Data Perspectives, Learning Paradigms and Challenges," arXiv preprint arXiv:2403.02990, March 2024. https://arxiv.org/abs/2403.02990

[18] H.F. Lin, M.X. Wu, X.L. Wang, Y.Q. Lin, L. Yu, L. Jiang, Q.S. Song, T.K. Kong, and W.X. Mu, "Few-shot biomedical NER empowered by LLMs-assisted data augmentation and multi-scale feature extraction," BioData Mining, Vol. 18, No. 28, April 2025. DOI: 10.1186/s13040-025-00443-y

[19] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," arXiv preprint arXiv:2010.11934, October 2020. https://arxiv.org/abs/2010.11934

[20] M. Junczys-Dowmunt, et al., "Marian: Fast Neural Machine Translation in C++," Proceedings of ACL 2018, System Demonstrations, Santa Fe, New Mexico, USA, pp. 120-125, July 2018. DOI: 10.18653/v1/P18-4020

[21] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Çelebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A.

www.kci.go.kr

Joulin, "Beyond English-Centric Multilingual Machine Translation," Journal of Machine Learning Research, Vol. 22, No. 107, pp. 1-48, 2021. DOI: 10.48550/arXiv.2010.11125

[22] D. Nozza, F. Bianchi, and D. Hovy, "What the [mask]? Making sense of language-specific BERT models," arXiv preprint arXiv:2003.02912, March 2020. https://arxiv.org/abs/2003.02912

[23] K. Saab et al., "Capabilities of Gemini Models in Medicine," arXiv preprint arXiv:2404.18416, April 2024. https://arxiv.org/abs/2404.18416

Authors

Tae-Hyeong Kwon received the B.S. in Computer Engineering from Hanyang University, Korea in 2017. He is currently an M.S. in the School of Computing, at Gachon University. He is interested in NLP, Chatbot, and Agent.

Dae-Ho Kim received the B.S. and M.S. degrees in Software from Gachon University, Korea, in 2017 and 2019, respectively. He is currently an Ph.D. student in the School of Computing, at Gachon University.

Se Young Kim received Ph.D. degree in Nursing from Seoul National University, Korea, in 2010. She is currently a Professor in the Department of Nursing, Changwon National University. She is interested in Nursing Surveillance Support system with AI & clinical decision making of nurses.

Ok-Ran Jeong received Ph.D. degrees in Computer Science and Engineering from Ewha Womans University, Korea, in 2005. She was a postdoctoral researcher at the University of Illinois at Urbana-Champaign, USA and Seoul National University, Korea. Dr. Jeong joined the faculty of the Department of Software Design & Management at Gachon University, Seongnam, Korea, in 2009. She is currently a Professor in the School of Computing, Gachon University. She is interested in big data mining, machine learning, deep learning and applications of artificial intelligence.

www.kci.go.kr