JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 5, pp. 21-30, May 2025
https://doi.org/10.9708/jksci.2025.30.05.021

Deep Learning based Automatic ICD Coding for Nursing Surveillance of Abdominal Surgery Patients

Dong-Hyeon Kim*, 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 paper, we propose an efficient dynamic workload balancing strategy which improves the performance of high-performance computing system. The key idea of this dynamic workload balancing strategy is to minimize execution time of each job and to maximize the system throughput by effectively using system resource such as CPU, memory. Also, this strategy dynamically allocates job by considering demanded memory size of executing job and workload status of each node. If an nodes, to another free nodes and reduces the waiting time and execution time of job by balancing workload of each node. Through simulation, we show that the proposed dynamic workload balancing strategy based on CPU, memory improves the performance of high-performance computing system compared to previous strategies.

▸ **Key words**: Nurse Surveillance, EMR, Automatic ICD Coding, Ensemble, Abdominal Surgery

[요 약]

간호감시는 임상 결과를 향상하는 중요한 역할을 하지만, 방대하고 복잡한 의료 데이터로 인해 진단명 분류에 어려움이 존재한다. 기존의 모델은 의사 중심의 데이터에 의존하거나 추가적인 기록을 요구하는 한계가 있었다. 본 연구는 이러한 한계를 해결하기 위해 EMR의 검사 데이터, 진단명 데이터, 간호기록 데이터만을 활용하여 간호감시 목적의 진단명 분류 모델을 제안한다. 복부수술 환자의 EMR 데이터를 바탕으로 환자 상태와 간호기록을 통합하여 KM-BERT 앙상블과 XGBoost를 결합한 모델을 구축하였다. 실험 결과, 제안한 모델은 단순 KM-BERT 및 기존 앙상블 모델보다 우수한 성능을 나타냈으며, 특히 KM-BERT 앙상블과 XGBoost를 결합한 모델이 가장 높은 정확도를 기록하였다. 본 연구는 EMR의 핵심 데이터만으로도 효과적인 진단명 분류가 가능함을 시사하며, 이는 간호감시의 효율성을 높이는 데 이바지할 수 있다. 앞으로의 연구에서는 보건의료 도메인 지식과 다양한 기법을 결합하여 모델 성능을 더욱 향상할 수 있을 것으로 기대된다.

▸ **주제어**: 간호감시, EMR, 진단명 분류, 앙상블, 복부수술

• First Author: Dong-Hyeon Kim, Corresponding Author: Se-Young Kim, Ok-Ran Jeong
*Dong-Hyeon Kim (eastlighting1@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. 02, Revised: 2025. 05. 18, Accepted: 2025. 05. 19.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

의료 현장에서 일하는 전문직에 대하여 크게 3가지 직무로 나눌 수 있다. 첫 번째는 치료(cure)에 관한 직무가 있다. 대표적으로 의사는 환자의 질병을 진단하고 낫게 하는 치료를 수행하고, 신체의 교정과 재활을 위한 물리요법적 치료를 하는 물리치료사나 신체적·정신적 기능장애를 회복시키기 위한 작업요법적 치료를 하는 작업치료사를 비롯한 몇 의료기사들이 참여하며 수행한다. 두 번째는 돌봄(cure)에 관한 직무가 있다. 대표적으로 간호사는 대상자의 건강을 사정하고 회복과 치료를 돕는 돌봄을 수행한다. 마지막은 검사(examination)에 대한 직무가 있다. 이 직무는 치료와 돌봄의 과정에서 이루어지는 검사를 진행하거나 그 검사 및 활동에 사용되는 기기를 관리하며, 임상병리사나 방사선사와 같은 일부 의료기사가 이 직무를 수행한다. 당연하게도 모든 의료인과 의료기사들의 행위가 중요하지만, 그중에서도 '의사가 없으면 진료과를 폐쇄하지만, 간호사가 없으면 병동 전체를 폐쇄해야 한다'[1]는 말이 있을 정도로 간호사가 진행하는 간호업무나 진료 보조 업무의 중요성은 상당하다. 빠르게 진화하는 의료 환경에서 효과적으로 진행되는 간호감시는 환자의 안전을 보장하고 임상 결과를 개선하는 데 중요한 역할을 한다. 그러나 수많은 양의 환자 정보와 의료 데이터의 복잡한 구조는 환자의 진단 및 간호사정을 시의적절하고 정확하게 분류하는 데 상당히 어려움을 주게 된다.

의료환경이 가지는 이러한 어려움을 해결하기 위해 딥러닝을 비롯한 첨단 기술들이 보조 수단으로서 등장했다. 데이터의 크기에 덜 구애를 받으면서 패턴의 식별과 의미있는 결과물의 도출을 이루어내는 능력을 갖춘 딥러닝 모델은 의료 분야에서 작동하였을 때 정확성과 효율성을 향상하는 것에도 좋은 성능을 보여주고 있다. 간호감시와 진단명 분류 태스크에서도 이러한 모델을 활용하면서 의료제공자의 인지적 부담을 덜어주고 위급한 상태를 조기에 감지하여 환자 치료를 개선할 수 있다. 건강보험에서 제공하는 2023 주요수술통계에 따르면, 34개의 주요 수술 중 충수절제술·담낭절제술·위절제술·간절제술·간색전술 등 주요 5개 복부 수술을 받는 환자가 전체의 13.2%를 차지하고 있다. 이처럼 많은 환자들이 복부 수술을 받고 있으며, 이렇게 많은 횟수의 수술 속에서 딥러닝 기반의 간호감시를 통해 체계적으로 간호중재를 진행하는 것이 중요해지고 있다. 그러나, 현재 제시되고 있는 딥러닝 모델들은 판독기록문과 같은 의사 중심의 데이터를 사용하고 있거나[2], EMR 데이터를 사용하더라도 온톨로지[3], 퇴원기록문[4] 등 별도의 기록을 추가로 요구하고 있다.

본 논문에서는 EMR에 포함된 검사 데이터와 간호기록 데이터만을 활용하여 의사의 치료 이후 작성되는 문서 없이도 진단명 코드를 예측하는 딥러닝 모델을 제안한다. 이 모델은 간호사의 간호사정과 간호진단 사이에서 간호사가 환자의 상태를 평가할 수 있는 정보를 제공한다. 이 모델에서 사용하는 매개체로는 EMR을 비롯한 다양한 직간접적 의료 시스템과의 상호운용성을 높일 수 있으며 판단에 대한 객관적 근거로서 추적가능성을 높일 수 있는 ICD 코드를 활용하였다. 본 연구는 질병의 예방·진단·치료를 담당하는 의학에서 병리 기전 등을 규명하기 위해 이루어지는 진단에 대한 예측을 활용하여 질병·치료·환경에 대한 생리·심리·사회적 반응을 파악해 간호중재를 위한 목표를 세우는데 기여한다. 또한 모델에 앙상블(Ensemble)과 XGBoost를 적용하여 모델의 예측 능력을 높이고, 그 두 알고리즘의 차원 차이를 보완하기 위해 PCA를 적용하는 방법을 제안한다.

II. Related Works

1. Nurse Surveillance

간호중재분류체계(NIC; Nursing Intervention Classification)에서 간호감시는 임상적 의사결정을 위해서 목적적이고 지속적으로 환자의 자료를 수집, 해석, 합성하는 것으로 정의한다[5]. 간호사는 환자와 상호작용하고 환자의 신체와 여러 지표, 간호중재에 대한 반응 등 다양하게 수집된 자료를 평가하며 환자의 위험을 구분한다[6]. 일반적으로 간호의 과정은 간호사정, 간호진단, 간호계획, 간호수행, 간호평가 5개의 세부 단계를 가지고 있다고 표현한다. 간호사정(Nursing Assessment)은 환자의 건강 상태와 관련 정보를 수집하고 분석하며, 간호진단(Nursing Diagnosis)은 환자의 건강 문제에 대한 명확한 정의를 제공한다. 간호계획(Nursing Planning)에서는 간호진단에 대하여 목표와 기대 결과를 설정하고 그에 맞추어 간호 행위를 선택하며, 이를 간호수행(Nursing Implementation)을 통해 실제로 수행한다. 간호수행이 끝나면 간호평가(Nursing Evaluation) 단계에서 효과를 평가하고, 필요한 경우 계획을 수정한다. 이를 통해 간호감시는 위해사건을 감소시키고 환자 안전을 증진하게 된다[7,34]. Eindhoven의 모형에서는 간호감시를 '환자를 위해 사건으로부터 보호하는 방어기전'으로 설명하고 있고[8], 이는 간호사가 위험한 상황을 알아채고 중단시키는 데 기여한다[9]. 간호감

시는 사회적 변화, EMR의 발전, 진단 및 추적 장치의 진보 등에 따라 환자의 건강과 안전을 위한 개념으로 확장되고 있다. 간호감시는 간호사의 전문성, 경험, 패턴인식, 직관, 지식 등 개인적 특성과 인력배치의 적절성, 스킬 믹스, 정보와 자원의 가용성, 전문직 실무환경 등 조직적 특성으로부터 영향을 받으며[10], 간호감시 교육, 시뮬레이션 훈련, 직관, PEWS, 상황인식, 임상적 지식, 경험이 그들의 역량에 영향을 주었다[11].

간호사가 간호감시 역량을 발휘하기 위해서는 최신 임상 지식과 해당 분야에서 오랜 근무 기간과 풍부한 경험을 쌓는 것이 필요하지만, 현실적인 측면에서 환자가 간호사가 성장할 때까지 언제까지고 기다릴 수는 없으며, 아무리 숙련된 간호사라고 하더라도 충분한 간호인력이 배치되지 않으면 간호감시를 효과적으로 수행하지 못한다[12]. 이를 해결하기 위해서, WHO는 활력징후와 검사 데이터를 모니터링하고 환자의 상태를 조기에 감지하여 처치할 수 있도록 돕는 EWS(조기경보시스템)을 제안했으며[13] 이는 부정적인 건강 결과를 예방하기 위한 중요한 요소가 되었다. 대표적인 EWS 시스템 중 하나인 VIDA[14]는 환자 데이터를 기반으로 등급을 나누어 감시에 대한 경고를 수행하고 임상적인 권고를 제시한 결과 VIDA를 사용한 병동의 COVID-19 환자 사망률을 다른 병동보다 낮출 수 있었다. 국내 역시 당뇨병학회에서 웹 기반 CDSS인 EGDM을 개발 및 보급하고 있다.

2. EMR

간호감시 과정에서 이루어지는 다양한 상호작용과 기록은 간호기록의 형태로 다양한 진료, 검사 기록들과 함께 EMR에 저장된다. EMR(Electronic Medical Record)은 환자와 인구의 건강 정보에 대해 디지털 형식으로 저장한 문서이다. EMR은 1990년대 개인처방전달시스템(OCS)이 등장한 이후로, 90년대 후반 영상저장전송시스템(PACS)이 등장하고 2000년대에 의료정보화가 이루어짐에 따라 환자 정보의 입력, 저장, 정보 교환 등의 시스템이 추가된 EMR이 구축되기까지 오랫동안 발전해 왔다. EMR의 도입은 환자 정보 획득을 위한 반복적인 작업을 줄여 병원의 재정적 손실과 시간 낭비를 사전에 방지할 수 있도록 한다.[15]

특히 EMR 시스템의 도입은 간호사에게도 큰 도움이 되었다. 초기 간호업무영역의 처방전달시스템은 몇몇 간호사들만 접근하였기 때문에 많은 활용을 하기가 어려웠지만, 의료정보기술의 발전에 따라 최근에는 환자의 사정 및 진단, 계획, 의사 처방 실행, 간호처치의 수가 입력, 투약, 의사를 향한 환자 상태 보고, 식이 입력, 전동 전실, 퇴원관리, 각종 물품 청구 및 관리에 이르기까지 다양한 영역에서 사용되고 있다. 간호사는 의사, 의료기사, 행정직 등 전문적이고 다양한 직종의 종사자들 간의 협업을 통해 환자들에게 의료서비스를 제공하는 간호업무를 수행하고, 환자들을 접하는 최일선에 있기 때문에 간호업무의 효율성 향상은 병원의 전반적인 업무 효율성과 직결되는 매우 중요한 부분이기도 하다[16].

EMR 시스템이 가지고 있는 의료 정보는 실제 환자의 정보를 전문 의료 인력이 정제한 데이터라는 점과 잘 구조화되어 있다는 점 덕분에 의료 도메인에서 AI를 개발하거나 데이터 분석을 진행할 때 데이터셋으로 적지 않게 활용되고 있다. 이도형 외 2인(2023)은 임상의가 다양한 검사와 신체 계측 정보, 안압 검사 등이 기록된 EMR 자료를 활용한다는 점에서 착안하여 안저사진과 EMR 데이터를 활용하여 녹내장을 예측하는 CNN+XGBoost 기반의 멀티모달 모델을 구현하였으며[17], 정진형 외 4인(2022)은 고령화 사회 진입 및 소득수준의 향상에 따른 스마트 헬스케어의 차원에서 접근해서 사용자가 실시간으로 EMR 속 환자 데이터를 볼 수 있는 아키텍처를 구성하여 간호사의 불안 요소를 줄일 수 있도록 하였다.[18] 김도원 외 5인(2022)은 욕창을 해결하기 위한 인공지능 모델에서 사용하는 데이터셋으로서의 EMR 데이터를 전처리하기 위한 방법으로서 시계열 데이터를 입실 후 사건기록까지의 길이로 변환하고 빈값을 채우는 방식을 제안하였다[19].

3. Automatic ICD Coding

Automatic ICD Coding은 의료 분야의 딥러닝 과제(task)의 일종으로, 임상 기록을 기반으로 ICD 코드를 최대한 정확하게 추측하는 것을 목표로 한다. ICD는 International Classification of Diseases의 준말로, 전 세계적으로 역학과 건강 관리에서 사용되는 분류 체계다. 임상 기록에는 환자의 전체 입원 동안 정확하게 무슨 일이 일어나는지 다양한 정보를 포함하고 있지만, 일반적으로 길고 스키마를 엄격하게 따르지도 않으며 철자 오류가 가득한 경우도 많다. 또한, 최근 발표된 ICD-11은 약 1만 7천 개의 코드와 12만 개가 넘는 용어를 포함하고 있기에[20] 매우 많은 라벨들을 가지고 분류를 진행해야 하는 'Extreme Multi-Label Classification'과제의 대표적인 예시이기도 하다.

Shurui(2022)는 임상 기록의 특정 부분이 코드를 할당하는 데 더 중요한 역할을 한다는 점, 코드 설명과 임상 문서의 이질성이 존재한다는 점 등을 고려하여 섹션별 임베딩을 통해 다층적 표현을 학습하는 Discourse Net과 문

서 간 표현 차이를 줄이기 위해 임베딩을 조절하는 Reconciled Embedding을 결합하는 방법을 제안하였다.[21] Tong(2021)은 학습 데이터의 불균형으로 인해 나타나는 long-tail 문제와 불필요하거나 중복된 정보로 인해 생기는 노이즈 문제를 해결하기 위해 코드 간 관계를 학습하는 공유 표현 네트워크를 도입하고 자가 증류 기법을 활용하여 중요한 정보에 집중할 수 있도록 유도하는 방법을 제안하였다[22].

4. Ensemble

앙상블(Ensemble)은 2개 이상의 모델을 결합하여 하나의 통합 모델을 구축하는 기법이다. 앙상블 기법은 의사결정 트리 혹은 인공신경망을 기반으로 생성된 기본 학습자(base learner)로 둔다. 앙상블을 구성하는 기본 학습자가 동일한 학습 알고리즘을 사용할 경우에는 이것을 동질적인 기본 학습자(homogeneous base learners)라고 하고, 동질적인 기본 학습자는 동질적 앙상블(homogeneous ensemble)을 생성한다. 만약 각 학습자가 동일한 학습 알고리즘을 사용하지 않는 경우에는 이질적 앙상블(heterogeneous ensemble)을 생성하며, 이때 학습자를 개별 학습자(individual learner) 혹은 구성(component learner)라고 하기도 한다. 앙상블 기법은 개별 기본 학습자보다 성능이 훨씬 뛰어난 경우가 많기에, 약한 학습자인 기본 학습자를 강한 학습자로 변환하는 데 큰 도움이 된다[23].

앙상블의 하위 기법으로는 대표적으로 Bagging, Boosting, Stacking이 있으며 이 기법들에 대한 응용으로 voting이나 blending 같은 것들이 제시되었다. Bagging[24]은 Bootstrap Aggregating의 약자로, 하나의 기본 학습자를 여러 버전으로 생성한 뒤 이를 활용하여 집계된 예측기를 얻는 방법이다. 기본 학습자의 예측 결과가 숫자인 경우에는 평균값을 사용하며, 분류기로 사용되는 경우에는 최다 득표 방식(plurality vote)를 사용한다. Voting은 서로 다른 기본 학습자를 사용하여 Bagging을 수행한다.

Boosting[25]은 편향을 개선하기 위해 제안된 모델로, 이전 모델이 잘못 예측한 부분을 개선하기 위해서 가중치를 부여하는 방식으로 분포를 조정한다. Boosting 기법은 크게 Adaptive Boosting 방식과 Gradient Boosting 방식이 있다. Adaptive Boosting 방식은 약한 학습기를 순차적으로 학습하며 틀린 값에 가중치를 부여하는 구조로 AdaBoost[26]에서 제안되었으며, 이외에도 다중 분류 문제로 확장한 SAMMA[27], 손실 함수를 로지스틱 손실로 바꾸는 LogitBoost[28] 등이 있다. Gradient Boosting 방식은 그래디언트 변화를 통해 오차를 줄이는 것에 집중하는 구조로 GBM[29]에서 제안되었으며, 이외에도 여러 최적화 기법을 추가한 XGBoost[30], 리프 중심의 트리 분할 방식을 사용하여 대용량 및 고차원 데이터에 최적화한 LightGBM[31], 범주형 데이터에 최적화한 CatBoost[32], 불확실성에 대한 정보를 제공하는 NGBoost[33] 등이 있다.

Stacking[34]은 메타 학습 기법으로 제안된 모델로, 여러 개의 기본 학습자를 학습시키고, 그 결과를 데이터셋으로 사용하여 메타 모델을 통해 최종 예측을 만드는 방식이다. Blending은 머신러닝 대회에서 stacking을 실용적으로 사용하기 위해 수정하는 과정에서 발견된 것으로, 단순히 검증값만을 활용해서 최종 예측을 진행하게 된다. Blending은 구현 가능성과 계산 비용이 적다는 단점이 있지만 과적합에 취약하고 성능 향상이 제한적일 수 있다는 단점이 있다.

III. Methodology

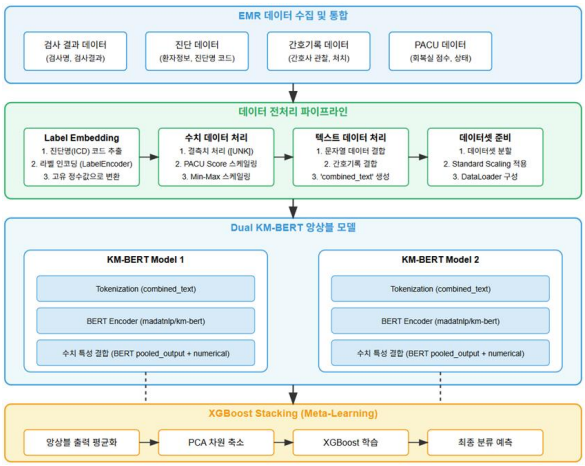

Fig. 1. System Architecture

Fig. 1.은 진단명 분류를 진행하기 위하여 XGBoost가 적용된 KM-BERT 앙상블 모델의 아키텍처이다. 이 모델은 정제를 위하여 전처리된 데이터를 입력으로 받으며, 그 이후에는 4가지 모듈을 통해 데이터를 처리한다.

첫번째 모듈인 Label Embedding Module은 범주형 데이터인 ICD 코드를 모델이 이해할 수 있는 숫자형으로 변환하는 역할을 수행한다. 이 모듈에서는 전체 라벨에 대하여 고유한 값들만 모은 후, 이를 라벨 인코딩을 해줄 인코더에 학습하고 실제로 적용한다. 두번째 모듈인 Feature Embedding Module은 모델의 훈련, 검증 및 테스트 데이터로 사용할 특성들을 적절하게 변환하고 결합하기 위한

Table 1. Some of Integrated Data

Patient ID	Diagnosis Code	Test Name	Test Result	Nursing Note	OR Nursing Note	Aggregated PACU Score
1	K81	WBC [응]	1.8	(지금)부터 …	지켜보자고 함 …	1.0
1	K81	RBC [응]	5.54	(지금)부터 …	지켜보자고 함 …	1.0
2	K81	Hb [응]	11.4	상복부 식사 후 …	NRS (3)점 …	1.0
2	K81	HCV Ab	Neg/	상복부 식사 후 …	NRS (3)점 …	1.0

임베딩을 진행하는 역할을 수행한다. 먼저 PACU Score과 같이 서열 척도인 데이터가 있는 경우에는 최소값과 최대값을 사용하여 Min-Max Scaling을 진행하며, 그 값이 비어 있는 경우에는 [UNK] 토큰을 통해 처리하고 최소값보다 1 낮은 값으로 고려한다. 이후에는 수치형 데이터와 문자형 데이터를 나누어 처리한다. 수치형 데이터는 Standard Scaling을 통해 표준 정규분포화하고, 문자형 데이터는 결합하여 dataloader를 생성한다. 세번째 모듈인 Dual KM-BERT Ensemble에서는 데이터를 2개의 KM-BERT에 각각 넣고 training을 진행한다. 이후 다음 모듈인 XGBoost Classifier을 통해 Stacking을 진행하게 된다. XGBoost는 다른 부스팅 알고리즘에 비해 계산이 경제적이며, 불균형한 클래스에 대응력이 있다는 장점이 있어 채택하였다.

1. Data and Preprocessing

본 연구에서는 Table 1과 같이 EMR 데이터 중 검사결과·IO·BST·활력징후 등의 환자 상태를 보여주는 데이터, 환자의 진단명 정보를 담고 있는 환자정보 데이터, 그리고 수술의 경과나 간호사정·간호중재 등의 정보를 담고 있는 간호기록·회복실 기록 등을 사용하였다. IO 데이터는 섭취와 배설에 대한 정보를 담고 있는 데이터로 식사·수액 등에 대한 경구/비경구 섭취 정보, 혈액·소변·대변·구토 등에 대한 배출 정보를 담고 있다. BST는 Blood Sugar Test의 준말로, 혈당에 대한 정보와 측정 당시에 혈당에 영향을 줄 수 있는 의약·식단 정보를 담고 있다. 활력징후는 신체의 생명에 관한 기능 정도를 판단할 수 있는 지표들을 담고 있는 데이터로, 체온·맥박·호흡·혈압·산소포화도 등에 대한 측정치를 담고 있다. 검사결과 데이터에서는 그외 검사에 대한 데이터를 담고 있다. 환자 정보 데이터에서는 나이·성별 등 환자 자체에 대한 정보, 진료과·진단명 등 입원 관련 정보와 마취·수술과·수술 중 행위 등 수술 관련 정보를 통합적으로 담고 있다. 간호기록은 간호를 위한 활동 중에 생기는 여러 정보들을 기록한 것으로, 대표적으로 DAR이라는 개념을 사용하여 작성하기도 한다. 또한 회복실 환자에 대해 PACU 점수에 대한 기록을 정리하기도 한다. PACU는 Post Anesthesia Care Unit의 준말로, 한국에서는 '(마취)회복실'이라고 번역되기도 한다. PACU 점수는 간호사들이 환자들의 회복 정도를 판단하기 위한 기준으로서 사용되는데, Activity(활동성), Respiration(호흡), Circulation(순환), Consciousness(의식). Skin Color (피부 색) 등을 지표로 사용한다. 이러한 정보들은 실시간으로 작성되거나 짧은 주기의 배치(batch)를 가지는 형태로 작성되기에 환자의 상태에 대한 정보로서 적합하고 간호사가 간호 행위를 하는 시점에서 빠르게 확인이 가능하다는 공통점을 가지고 있다.

다만 이러한 정보를 바로 가져다 쓰기에는 feature의 수가 너무 많고 출처가 되는 파일 별로 데이터의 형태가 다르기에 전처리가 필요하다. 먼저 동일한 목적을 가지고 있으며 구조적으로 통합이 가능한 검사결과·IO·BST·활력징후 데이터를 통합하였다. 검사결과 데이터를 바탕으로 IO·BST·활력징후 데이터를 '검사 결과'와 '검사명'를 작성하는 방식으로 정리하였다. PACU 점수의 경우에는 각 지표 별, 경과 시간 별로 열이 존재하고 있다. 일반적으로 PACU 점수는 특정 경과 시간에 측정된 PACU 지표의 각 점수들 간 총합으로 사용된다는 점에서 착안하여, 회복실 퇴원 시점에서의 총합 점수에서 회복실 입원 시점에서의 총합 점수를 빼는 방식으로 집계를 수행하였다. 이후에는 환자 별로 부여된 ID를 중심으로 inner join 방식으로 통합하여 최종 입력 데이터를 구성하였다.

2. Ensemble Learning

본 연구에서는 스태킹 방식으로 앙상블을 수행하였다. Fig. 2.는 모델에서 사용된 앙상블 학습의 구조를 간략화한 것이다. 먼저 2개의 KM-BERT 모델에 대하여 각각 학습을 진행한다. KM-BERT 모델은 한국어 기반의 자연어처리 모델인 KR-BERT를 베이스 모델로 하여 2만 5986개의 정리된 의학 용어 집합을 바탕으로 의학과 관련된 약 6백마내의 문장과 약 1억 1600만개의 단어를 학습한 모델이다. 의료 분야의 용어는 'common cold - acute infective rhinitis' 처럼 같은 의미를 가진 일상 용어와는 다른 단어를 사용하는 경우가 많아 임베딩과 처리를 용이하게 하기 위해 사용하였

다. 두 모델의 학습이 끝나면 그 결과값에 대해 평균을 구하며, 그 결과에 대해 XGBoost를 메타 모델로 활용하여 학습을 진행한다. 다만 KM-BERT의 출력 차원은 클래스 수에 의존적이며 진단명 분류 과제는 매우 많은 클래스를 전제하기 때문에 굉장히 고차원인데 반해, XGBoost는 내부적으로 의사결정 트리를 사용하는 특성 상 고차원에 다소 취약할 수 있다. 따라서 평균이 된 결과 값을 XGBoost에 넣기 전에 PCA를 통하여 차원을 축소한다.

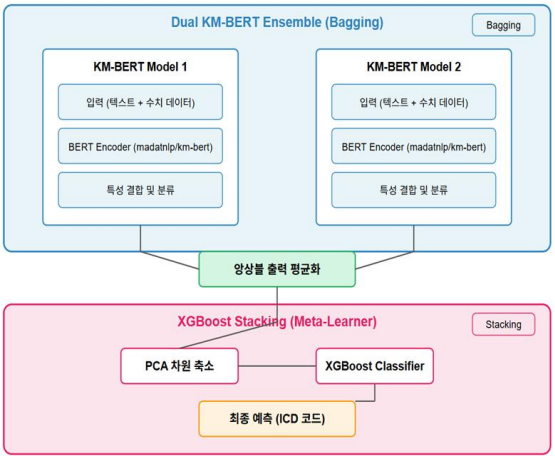

Fig. 2. Ensemble Architecture

이 모델에서 제안하는 앙상블 구조를 상세하게 살펴보면 큰 단위에서 Bagging 기법과 Boosting 기법이 적용되었으며 그 두 기법을 Stacking 기법을 통해 연결하고 있음을 알 수 있다. Dual KM-BERT가 출력한 raw logits을 샘플 별로 평균해 행렬을 만들고, 여기에 PCA를 적용해 축소된 벡터를 얻는다. 이 평균 로짓 벡터가 그대로 XGBoost의 feature matrix로 들어가 최종 예측을 담당한다. 동일한 두 KM-BERT의 평균 결과값을 구하는 방식은 Bagging 기법에 해당하며, XGBoost는 대표적인 Boosting 기법의 모델 중 하나이다. 그리고 Bagging을 통해 생성된 앙상블 출력은 XGBoost라는 Boosting 기법을 메타 모델로 하는 Stacking 기법의 입력 데이터가 된다. 이렇게 다양한 앙상블 기법을 적용하게 되면 특정 기법이 데이터의 잡음을 학습하는 것을 방지하여 일반화 성능을 높일 수 있고, 단순히 하나의 기법이 포착하기 어려운 패턴을 더 잘 학습할 수 있다.

IV. Experiment

이 장에서는 제안하는 모델의 실험에 사용한 데이터셋과 실험 환경, 평가 지표, 그리고 결과에 대하여 설명한다.

1. Dataset

모델의 학습 및 평가에 사용된 데이터는 경남 지역 A 상급종합병원의 의무기록 중 일부이다. 2018년 10월 1일부터 2023년 9월 30일까지 일반외과 병동에 입원하여 위절제술, 간절제술 등 복부수술을 받은 환자를 대상으로 수집하였으며, 위절제술이나 간절제술 이외에 다른 장기 수술을 받았거나 안전사고가 발생한 환자 등은 배제하였다. 수집된 환자수는 8587명이며, S 대학병원의 IRB 승인과 데이터심의위원회 사용 승인 절차를 거친 후 병원정보시스템과 EMR 담당자와의 협의를 거친 후 사용하였다. 본 연구에서는 scikit-learn의 train_test_split 메소드를 통해 훈련, 검증, 실험에 각각 40%, 30%, 30% 비율로 사용하였으며 stratify 옵션을 통해 각 라벨 별 비율이 똑같이 유지될 수 있도록 하였다. 또한, Pytorch 라이브러리에서 제공하는 WeightedRandomSampler 메소드를 이용하여 라벨 별 비율이 유지되면서 샘플링을 진행할 수 있도록 하였다. train sampler에 대한 replacement 옵션은 True로, validation과 test에 대한 replacement 옵션은 False로 설정하였다. 이외에도 공통적으로 subprocess(num_workers)를 16개로 두었으며, 배치 크기는 16으로 하였다. 또한 앞서 언급하였듯이 Automatic ICD Coding는 Extreme Multi-Label Classification에 해당하는 작업 중 하나이므로, 데이터의 복잡성에 대한 이해도 필요하다. 본 연구에서 사용하는 데이터는 141개의 ICD 코드로 구성되어 있다.

Fig. 3.은 코드의 분포를 나타낸 것이다. 이외에도 가장 개수가 많은 코드는 833개이며 가장 개수가 적은 코드는 20개로, 클래스 불균형 비율이 약 41.8배로 높은 편이다. 분포 전체의 불균형성을 보기 위해 Gini index를 계산하였을 때는 0.62로, 꽤 불균형함을 알 수 있다.

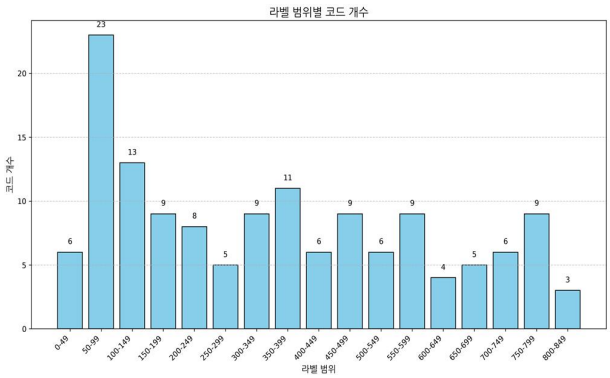

Fig. 3. Histogram for Label Count

2. Environment and Metrics

모델의 학습과 실험에서 사용한 컴퓨터의 프로세서는 AMD Ryzen Threadripper 1950X이며, 제공된 최대 메

Table 2. Experiment Result

Model	Accuracy	Weighted Precision	Weighted Precision	Weighted Precision
Single KMBERT	0.4368	0.3735	0.4368	0.3757
Single KMBERT+XGBoost	0.9197	0.9027	0.9197	0.9085
Single KMBERT + XGBoost + PCA	0.9234	0.9093	9.9234	0.9146
Double KMBERT	0.4383	0.4034	0.4383	0.3868
Double KMBERT+XGBoost	0.9232	0.9084	0.9232	0.9130
Double KMBERT + XGBoost + PCA	**0.9245**	**0.9107**	**0.9245**	**0.9157**

모리는 126GB이다. 사용한 GPU는 GeForce RTX 2080 SUPER이며, 제공된 최대 VRAM은 16GB이다. 운영체제는 Ubuntu 24.04.1 LTS이며, 최대 저장공간은 2TB이다.

모델은 머신러닝 라이브러리인 Pytorch와 데이터 분석 라이브러리인 Polars를 활용하여 구현하였으며, 데이터의 정합성을 위해 Dask 라이브러리를 사용하여 데이터클래스를 규정하였다. 모델의 학습에는 CPU와 GPU를 모두 사용하였다. 모델의 평가지표로는 Accuracy, Precision, Recall, F1 Score을 사용하였으며 라벨 간 분포가 고르지 않은 점을 감안하여 클래스 균형도에 영향을 받는 Precision, Recall, F1 Score에 대해서는 가중평균된 평가지표를 활용하였다.

Table 3. Hyperparameters

KM-BERT	max_length	512
	dropout_prob	0.30
	optimizer	Adam
	learning_rate	2e-5
	lr_scheduler	StepLR
	batch_size	16
PCA	n_components	0.95
XGBoost	objective	multi: softprob
	num_class	141
	max_depth	6
	early_stopping_rounds	10
	eval_metric	mlog loss

실험에서 활용한 주요 하이퍼파라미터는 Table 3과 같다. 작성하지 않은 경우, 기본값을 사용하였다.

3. Result

실험 결과는 Table 2와 같다. 사용한 데이터셋에 대하여 단순 KM-BERT 모델과 KM-BERT Ensemble 모델로 나누어 실험하였으며, 각 모델에 대하여 XGBoost 도입 여부와 PCA 도입 여부에 따라 나누어 8개의 경우의 수가 있지만 PCA의 도입 목적 상 XGBoost가 도입되지 않은 경우 PCA도 도입하지 않아 총 6가지의 실험 모델을 구성하였다. 수치와 텍스트가 모두 있는 데이터의 특성, 연구의 목적, 그리고 기존 연구들이 신경망 이전 머신러닝 모델 위주로 연구되었음을 고려하여 성능의 안정성을 보기 위해 KM-BERT를 중심으로 ablation study 방식의 실험을 택하였다. 각 모델의 epoch는 10으로 제한하였으며 early-stopping 기법을 사용하였다. 제안 모델과 비교 모델 모두 early-stopping으로 epoch 10 이전에 종료되었다. 훈련의 경우에는 검증 단계로 넘어가기까지 약 1244초가 소요되었다. Train과 validation에서 사용한 sample의 수는 1.6만개, test에서 사용한 sample의 수는 5만개로 하였다. 또한, BERT 기반 미세조정 모델인 KM-BERT의 임베딩으로 구성된 feature space에 대하여 전체 분산의 95%를 보존하는 최소한의 주성분을 선택하였다. PCA fitting 작업은 학습 데이터에서만 수행하였으며, 검증과 테스트에서는 동일한 변환을 일관되게 적용하여 데이터 누수를 방지하였다. PCA와 XGBoost를 모두 적용한 경우에는 'feature scaling - PCA 적용 - XGBoost'와 같이 파이프라인을 구성하였고, 하나 이상 제외된 경우에는 해당하는 단계를 제외하였다.

Table 2의 데이터를 보면 앙상블 자체의 도입 여부, PCA 도입 여부, XGBoost 도입 여부 각각에 대하여 모두 긍정적인 성과를 보여주고 있는 것을 볼 수 있다. 특히, 이 모든 것이 적용된 'Double KM-BERT + XGBoost + PCA' 모델의 경우에서는 Accuracy, Precision, Recall, F1 Score 등 모든 지표에서 가장 성능이 높은 것을 확인할 수 있다. Table 4는 더 자세한 비교를 위해 앙상블 유무에 따른 배치를 한 것이다. KM-BERT 모델만 단독으로 있는 경우, XGBoost를 통해 Stacking이 되는 경우, PCA까지 전부 적용한 경우에서 모두 앙상블이 적용된 상황에서 더 높은 정확도를 보이는 것을 알 수 있다.

Table 4. Comparison of Accuracy Based on Ensemble Usage

Model Type	Double	Single
KM-BERT+XGBoost+PCA	**0.9245**	0.9234
KM-BERT+XGBoost	**0.9232**	0.9197
KM-BERT	**0.4383**	0.4368

Table 5. Comparison of Accuracy Based on XGBoost Usage

Model Type	Apply	Not Applied
Single KM-BERT	**0.9197**	0.4368
Double KM-BERT	**0.9232**	0.4383

Table 5 또한 더 자세한 비교를 위해 XGBoost의 적용 여부에 따른 배치를 한 것이다. 만약 이 비교에서 PCA를 포함할 경우 정확도에 영향을 주는 부분이 PCA 때문인지 아니면 XGBoost 때문인지 구분할 수 없기 때문에, 이 비교에서는 PCA가 적용되지 않은 경우를 기준으로 두 타입을 선정하였다. 이 비교에서는 적용 여부에 따른 정확도의 차이가 약 2배 가까이 보이고 있다. 이는 XGBoost와 이를 위시하는 Stacking 기법이 모델의 특성에 대한 학습과 분류하는 능력을 강화하는데 주요한 능력을 하고 있다고 볼 수 있다.

Table 6. Comparison of Accuracy Based on PCA Usage

Model Type	Apply	Not Applied
Single KM-BERT	0.9234	0.9197
Double KM-BERT	0.9245	0.9232

마지막으로, Table 6는 PCA의 적용 여부에 따른 배치를 한 것이다. 이때도 적용한 경우의 정확도가 적용하지 않은 경우의 정확도보다 소폭 높은 것을 볼 수 있다. 이는 PCA가 데이터의 차원을 축소하여 노이즈를 줄이는 부분이나 모델 학습의 효율성을 높이는 부분에서 긍정적인 역할을 하고 있다고 해석할 수 있다.

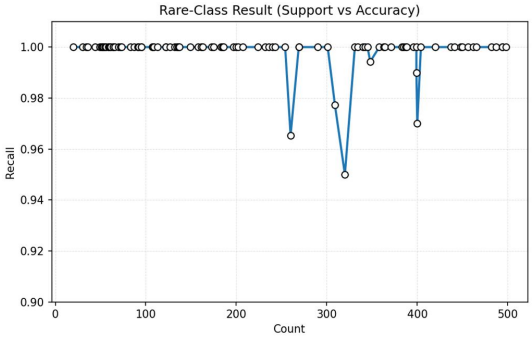

Fig. 4. Rare Class Result

Fig. 4.는 제안 모델이 비교적 개수가 적은 클래스에서도 잘 작동하는지 확인하기 위해 샘플을 기준으로 1% 이하의 개수를 가진 희소 클래스(rare class)에 대한 정확도를 나타낸 것이다. 클래스 별 정확도의 경우 recall과 같이 $\frac{TP}{TP+FN}$이 되므로 recall을 활용하였다. 희소 클래스의 경우에도 거의 대부분의 클래스에서 굉장히 높은 recall 값을 보여주고 있으며, 가장 낮은 값의 경우에도 95%의 확률로 예측하고 있다.

V. Conclusions

본 논문에서는 복부수술 환자의 간호감시를 위해 EMR 데이터 중 환자 진단, 환자 상태, 간호기록 데이터에 해당하는 데이터를 활용하여 KM-BERT 모델과 XGBoost를 이용한 진단명 분류 모델을 제안하였다. 제안한 모델은 단순 KM-BERT 모델이나 KM-BERT에 단순히 bagging을 적용한 모델과 비교하였을 때 월등한 성능을 보여준다는 것을 알 수 있었다. 본 연구는 복부수술 환자의 데이터를 활용한 간호감시 목적의 진단명 분류의 기초적인 모델을 구축하였다는 것에 의의가 있다. 트랜스포머 구조의 모델 아키텍처는 BERT 이후로도 XLNet, SpanBERT, RoBERTa 등 다양하게 나오고 있고, 부스팅 기법 역시 XGBoost 이외에도 Adaboost, LightGBM, CatBoost 등 다양하게 제시되고 있다. 또한, 본 연구에서 구현한 모델은 보건의료 도메인에 대한 지식보다는 기계학습과 딥러닝 중심의 지식을 활용하여 구현하였다는 한계점이 존재한다. 향후 연구에서는 보건의료 도메인에 대한 지식과 함께 이러한 기법들을 활용하여 더 발전된 모델을 구축할 수 있을 것으로 기대한다.

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] If there is no doctor, the department closes, but if there is no nurse, the entire ward closes., *https://www.medicaltimes.com/Main/News/NewsView.html?ID=1099215*

[2] Kyeongjin Ann, Yeonggul Jang, Youngtaek Hong, Sunghee Jung, Hackjoon Shim, and Hyuk-JaeChang, "Final Diagnosis Classificati on using Fine-tune BERT for Automatic Labeling," Journal of the Institute of Electronics and Information Engineers, Vol. 56, No. 12, pp. 92-98, December 2019. DOI: 10.5573/ieie.2019.56.12.92

[3] D. Bruness, M. Bay, C. Schulze, M. Guckert, and M. Minor, "A Hybrid AI-Based Method for ICD Classification of Medical Documents," Studies in Health Technology and Informatics, Vol. 305, pp. 1-4, June 2023. DOI: doi:10.3233/SHTI230408

[4] I. Aden, C. H. T. Child, and C. C. Reyes-Aldasoro, "International Classification of Diseases Prediction from MIMIIC-III Clinical Tex t Using Pre-Trained ClinicalBERT and NLP Deep Learning Model s Achieving State of the Art," Big Data and Cognitive Computing, Vol. 8, No. 5, p. 47, May 2024. DOI: 10.3390/bdcc8050047

[5] J. M. Dochterman, C. M. Wagner, H. K. Butcher, and G. M. Bulechek, *Nursing Interventions Classification (NIC)-E-Book*, Elsevier Health Sciences, 2018.

[6] S. Dresser, "The role of nursing surveillance in keeping patients safe," The Journal of Nursing Administration, Vol. 42, Issue 7/8, pp. 361-368, July/August 2012. DOI: 10.1097/NNA.0b013e31826 19377

[7] 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.

[8] T. W. van der Schaff, *Near-Miss Reporting in the Chemical Process Industry*, Ph.D. Thesis, Eindhoven University of Technology, The Netherlands, 1992.

[9] E. A. Henneman, A. Gawlinski, and K. K. Giuliano, "Surveillance: a strategy for improving patient safety in acute and critical care units," Critical Care Nurse, Vol. 32, No. 2, pp. e9-e18, April 2012. DOI: 10.4037/ccn2012166.

[10] M.-E. Juvé-Udina et al., "Surveillance nursing diagnoses, ongoing assessment and outcomes on in-patients who suffered a cardioresp iratory arrest," Revista da Escola de Enfermagem da USP, Vol. 51, p. e03286, March 2018. DOI: 10.1590/s1980-220x201700470 3286

[11] J. Thrasher, H. McNeely, and B. Adrian, "When nursing assertion stops: A qualitative study to examine the cultural barriers involved in escalation of care in a pediatric hospital," Critical Care Nursing Clinics of North America, Vol. 29, No. 2, pp. 167-176, June 2017. DOI: 10.1016/j.cnc.2017.01.004

[12] A. Kutney-Lee, E. T. Lake, and L. H. Aiken, "Development of the hospital nurse surveillance capacity profile," Research in Nursing & Health, Vol. 32, No. 2, pp. 217-228, April 2009. DOI: 10.1002/nur.20316.

[13] WHO, *Clinical management of severe acute respiratory infection when novel coronavirus (nCoV) infection is suspected*, 2020.

[14] J. Adamuz 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. 2, p. e041726, February 2021. DOI: 10.1136/bmjope n-2020-041726

[15] Choyeal Park, "A Study on the Recognition of Certification Syste m in EMR Certified Medical Institutions - Focusing on EMR Certification and System Functionality of Public Healthcare Instit utions," Korea Society of Health Service Management, Vol. 17, No. 4, pp. 1-14, December 2023. DOI: 10.12811/kshsm.2023.17. 4.001

[16] Kang, Jisook, Kim, Sunja, and Kim, Wonjeong, "The Autonomy, Nursing Performance based on the Awareness and satisfaction of EMR System for Nurses," Journal of the Korea Academia-Industrial cooperation Society, Vol. 16, No. 9, pp. 6061-6070, September 2015. DOI: 10.5762/KAIS.2015.16.9.6061

[17] DoHyeong Lee, Seong Jae Kim, and Sejong Oh, "A Multimodal Glaucoma Diagnosis Model based on Fundus Image and EMR Data," The Journal of Korean Institute of Information Technology, Vol. 22, No. 6, pp. 13-19, June 2024. DOI: 10.14801/jkiit.2024.22.6.13

[18] Jin-Hyoung Jeong, Jae-Hyun Jo, Seung-Hun Kim, Won-yeop Park, and Sang-Sik Lee, "A Study on the Development of Intravenous Injection Management Application for EMR System Interworking," Journal of Korea Institute of Information, Electronics, and Communication Technology, Vol. 15, No. 6, pp. 506-514, December 2022. DOI: 10.17661/jkiiect.2022.15.6.506

[19] Dowon Kim, Minkyu Kim, Yoon Kim, Seon-Sook Han, Jungwon Heo, and Hyun-Soo Choi, "Method of preventing Pressure Ulcer and EMR data preprocess," Journal of the Korea Society of Computer and Information , Vol. 27, No. 12, pp. 69-76, December 2022. DOI: 10.9708/jksci.2022.27.12.069

[20] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, CRC Press, 2012.

[21] L. Breiman, "Bagging predictors," Machine Learning, Vol. 24, pp. 123-140, August 1996. DOI: 10.1007/BF00058655

[22] R. E. Schapire, "The Strength of Weak Learnability," Machine Learning, Vol. 5, No. 2, pp. 197-227, June 1990. DOI: 10.1007/B F00116037

[23] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Proceedings of EuroCOLT 1995, pp. 23-37, Lecture Notes in Computer Scien ce, Vol. 904, Springer, Berlin, Heidelberg, January 2005. DOI: 10.1007/3-540-59119-2_166

[24] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class adaboost," Statistics and its Interface, Vol. 2, No. 3, pp. 349-360, 2009. DOI:

10.4310/SII.2009.V2.N3.A8

[25] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," The Annals of Statistics, Vol. 28, No. 2, pp. 337-407, April 2000. DOI: 10.1214/aos/1016218223

[26] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, Vol. 29, No. 5, pp. 1189-1232, October 2001. DOI: 10.1214/aos/1013203451

[27] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, New York, Unite States. August 2016. DOI: 10.1145/2939672.2939785

[28] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," Advances in Neural Information Processing Systems, pp. 3149-3152, California, Unite States, Vol. 30, December 2017.

[29] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," Advances in Neural Information Processing Systems, pp. 6639-6649, Louisiana, Unite States, Vol. 31, December 2018.

[30] T. Duan et al., "NGBoost: Natural Gradient Boosting for Probabilistic Prediction," Proceedings of the International Conference on Machine Learning, pp. 2690-2700, Virtual, July 2019. DOI: https://doi.org/10.48550/arXiv.1910.03225

[31] D. H. Wolpert, "Stacked generalization," Neural Networks, Vol. 5, No. 2, pp. 241-259, 1992. DOI: 10.1016/S0893-6080(05)80023-1

[32] WHO, "ICD-11 2022 Release," https://www.who.int/news/item/11-02-2022-icd-11-2022-release

[33] S. Zhang, B. Zhang, F. Zhang, B. Sang, and W. Yang, "Automatic ICD coding exploiting discourse structure and reconciled code embeddings," Proceedings of the 29th International Conference on Computational Linguistics, pp. 2883-2891, Gyeongju, Republic of Korea, October 2022. DOI: 10.48448/hnw4-6c74

[34] S. Y. Kim and M. K Cho, “Concept analysis of nursing surveillance using a hybrid model,” Healthcare, Vol. 11. No. 11., pp. 1613, May 2023. DOI: 10.3390/healthcare11111613

Authors

Dong-Hyeon Kim received the B.S. degree in Software from Gachon University, Korea, in 2024. He is currently an M.S. student in the School of Computing, at Gachon University. He is interested in Agentic AI, Data Engineering, Data-centric AI and NLP.

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.