Check for updates

Journal of KIIT. Vol. 22, No. 2, pp. 43-52, Feb. 28, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.2.43

건설-제조 현장 작업자의 복합 질환 선별을 위한 건강 검진 데이터를 사용한 이진-관련성-기반 다중 레이블 분류

등덕호*, 이태림**, 변승혁***, 최문택****

Binary-Relevance-based Multi-Label Classification using Medical Examination Data to Screen for Complex Illnesses in Construction and Manufacturing Workers

Teh-Hao Teng*, Tae-Rim Lee**, Seung Hyeok Byeon***, and Mun-Taek Choi****

이 연구는 2023년도 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원에 의한 연구임('20015190')

요 약

본 연구의 목표는 제조 및 건설업 현장에서 근무하는 작업자들의 직군 별 건강 검진 데이터를 활용하여 건강 상태를 복합 질환 차원에서 예측하는 인공지능 모델의 가능성을 탐구한다. 한 사람이 여러 개의 질환에 해당 될 수 있어서 다중 레이블 분류를 위한 모델은 문제 변환의 대표적인 이진 관련성(BR)을 기반으로 개발하였다. 데이터 세트는 건설 및 제조 현장 작업자 601명을 대상으로 수집한 건강 검진 데이터 82개의 특징과 9개의 질환 레이블로 구성되어 있다. 다양한 분류 알고리즘들을 그리드 서치 기반의 하이퍼파라미터 튜닝을 통해서 학습시키면서 최적의 모델을 선택했다. 이는 건설·제조업 현장 작업자에게서 의심되는 복합 질환의 식별에 대규모 건강 검진 데이터를 사용해서 최초로 인공지능 모델을 적용한 연구라고 할 수 있다.

Abstract

The objective of this study is to explore the potential of building an artificial intelligent model with medical examination data from construction site workers to predict health status in terms of multiple illnesses. The data set, collected from 601 field workers, consists of 82 features and 9 illness labels. Given that an individual may be diagnosed with multiple illnesses, the AI model adopts a binary relevance-based multi-label classification approach. Various classification algorithms were trained through grid search-based hyperparameter tuning to select the optimal model. This is the first study to apply artificial intelligence models using large-scale medical examination data to identify suspected complex illnesses in construction and manufacturing workers.

Keywords

complex illnesses, health medical data, construction workers, binary relevance, multi-label classification

* 성균관대학교 지능형로봇학과 석사과정
- ORCID: https://orcid.org/0009-0002-8153-7901
** 성균관대학교 인공지능학과 석.박사통합과정
- ORCID: https://orcid.org/0000-0002-2676-5728
*** ㈜인피니티케어 R&D연구소 연구소장
- ORCID: https://orcid.org/0009-0005-1176-0894
**** 성균관대학교 지능형로봇학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-5025-8785

· Received: Jan. 16, 2024, Revised: Feb. 16, 2024, Accepted: Feb. 19, 2024
· Corresponding Author: Mun-Taek Choi
Sunkyunkwan University
Tel.: +82-31-290-4929, Email: mtchoi@skku.edu

Ⅰ. 서 론

건설 · 제조 현장의 근로자의 근로 환경과 건강 영향 등에 대한 다양한 연구가 발표되고, 직종에 따라 자주 발생되는 질환에 대한 연구도 활발하게 진행되고 있다. 하지만 기존의 연구는 건설 · 제조 근로자의 주요 질환 도출에는 의미가 있지만 질환 예방을 통한 근로 환경 개선 활동에는 한계가 있다[1]. 특히 산업 내 개인의 건강 관리를 통한 산업 재해 예방 및 업무 생산성 관리는 불가능하다. 건설 및 제조 현장 근로자의 특성에 맞는 건강 진단을 검토하고 근로자의 건강 보호 방안을 마련하는 것이 시급하다.

한편, 최근 인공지능 기술의 발달은 다양한 분야에 적용되고 있으며, 특히 의료 영역에서는 질병을 예측하고 진단하는 것을 넘어 의학적 해석이나 개인의 건강 관리 의료 서비스 제공까지 그 역할이 확대되었다[2]. 현재의 인공지능 영역은 가상 간호 보조, AI 기반 로봇 수술, 영상 해석, 의료 진단 및 의료 서비스로 요약할 수 있다[3]. 의료 진단 및 의료 서비스 영역은 개인 의무 기록을 수집 · 분석을 통해 새로운 의학 지식을 발견하거나 질병 예측을 위한 모델링 도구를 통해 업무에 적용할 수 있다.

앞서 언급한 문제점과 기술적 동향을 바탕으로 본 연구에서는 제조 · 건설 현장 근로자의 근무 중 발생 가능한 건강 위험 요인을 발견하고, 인공지능 기반의 건강 위험 요인 특징 예측이 가능한 산업용 의료 서비스의 가치를 탐구해 보고자 한다. 일반적으로 근무자들의 건강 상태는 다양한 질환을 복합적으로 갖고 있고 이에 영향을 미치는 요인도 매우 다양하기 때문에 인공지능 접근에서 이 문제는 다중 레이블(Multi-label) 분류 문제로 다루어야 한다[4].

다중 레이블 분류 문제에서는 하나의 데이터 인스턴스가 여러 개의 레이블을 가질 수 있다[5]. 기계 학습에서 다중 레이블 분류 문제에 대한 해결책으로는 문제 변환 방법과 알고리즘 적응 방법이 있다. 문제 변환 방법의 핵심 개념은 다중 레이블 문제를 단일 레이블 문제의 해결책으로 접근하는 것이다. 여기에는 이진 관련성(BR, Binary Relevance), label powerset, 쌍 별 비교를 통한 순위 지정 등이 포함된다. 알고리즘 적응 방법의 핵심 개념은 다중 레이블 문제를 처리하기 위해 단일 레이블 분류기를 수정하는 것이다. 여기에는 다중 레이블 의사 결정 트리, 다중 계층 신경망, 다중 레이블 K 최접근 이웃 등이 포함된다.

BR은 다중 레이블 분류 문제를 해결하는 직관적인 방법으로 각 레이블에 대해 별도의 이진 분류를 생성한다. 여러 연구에서 BR을 사용하여 다중 레이블 분류기를 개발하려고 시도했다. M. L. Zhang and L. Wu[6]는 BR을 사용하여 각 레이블에 대해 특정 특징을 사용했다. 이들은 다중 레이블 분류가 레이블 별 특징에 의해 더 나은 성능을 보인다는 것을 보여주었다. J. Huang et al.[7]은 상호 연관된 레이블 간에 레이블 별 특징을 공유하는 방법을 제안했다. 다른 다중 레이블 분류 방법과 비교했을 때 경쟁력 있는 성능을 보였다. E. Alvares-Cherman et al.[8]은 BR에서 레이블의 상관관계를 반영하지 못하는 한계를 극복하기 위해 노력했다. 이들은 분류자 체인 방식보다 더 나은 결과를 보였지만, 각 레이블에 대해 다양한 분류 알고리즘을 적용하지는 않았다.

본 연구는 대상자의 질환 예측 모델을 만들 때 한 사람이 여러 질환에 해당될 수 있어서 가장 기본이 되는 BR 방법을 기반으로 다중 레이블 기반 분류 문제로 다루고자 했다. 다양한 분류 알고리즘을 학습하고 그 성능을 비교하여 작업자 건강 상태 선별을 위한 최적의 모델을 선정하고자 했다. 이러한 연구는 건설 및 제조업 현장에서 흔히 발생하는 다양한 질환을 식별하고 예방하기 위한 새로운 방법을 제시하며, 작업자들의 건강과 안전을 더욱 향상 시키는데 기여할 것으로 기대된다. 이어지는 연구 과정에서 더 나은 모델과 접근 방식을 개발하여 현장 작업자들의 건강을 더욱 효과적으로 관리하는데 기여하고자 한다.

Ⅱ. 샘플 데이터

앞서 언급했던 작업을 위해 2021년 4월부터 2023년 12월까지 총 601명의 건설 · 제조 근로자들을 대상으로 한신메디피아에서 건강 검진을 실시하고, 개인 의무 기록을 수집하여 데이터 기반 인공지능을 활용한 분석을 수행하였다[9].

개인 의무 기록 활용을 위해 한신메디피아 기관 생명윤리위원회로부터 IRB 승인을 받았다. (IRB 승인 : HS23-3, 2023. 4. 12) 종합 건강 검진 데이터의 활용을 위해 의료 기관 내 ZeTTA PACS 계정을 통해 인공지능용 DB를 구축하였다. 이러한 건강 검진 데이터에는 총 82개의 검진 항목 특징 요소와 다중적 질환 유무를 나타내는 9개의 레이블로 구성되어 있다. 비만(Obesity), 간 질환(Liver disease), 고혈압(Hypertension), 당뇨(Diabetes), 신장 질환(Kidney disease), 이상지질혈증(Dyslipidemia), 순환기 질환(Cardiovascular disease), 근골격 불균형(Musculoskeletal imbalance), 스트레스(stress). 해당 질환 유무 데이터는 각 인원들이 9개의 질환 중에 어떤 것에 양성(1)혹은 음성(0)인 보여 준다.

표 1. 주요 특징 및 질환 레이블
Table 1. Features and disease labels

Term	Element
Main features	body mass index, blood pressure (diastolic/export), waist circumference, platelet count, obesity, weight, triglycerides, height, uric acid, creatinine, urinary specific gravity, T4, splenocytes, B/C ratio, HDL, BMD, CPK, HbA1c, TIBC, etc.
Illness labels	obesity, liver disease, hypertension, diabetes, kidney disease, dyslipidemia, cardiovascular disease, musculoskeletal imbalance, stress

III. 다중 레이블 질환 분류 모델

3.1 전처리

전처리 과정은 건강 검진 데이터에 대한 효과적인 머신러닝 모델 구축을 위한 필수적인 단계이며, 데이터의 정제와 표준화는 모델의 성능과 일반화 능력에 직접적인 영향을 미친다. 건강 검진 기반 검사 측정 특징 및 레이블 데이터 세트에서 목표로 하고자 하는 질환 열을 타겟 변수로 설정하고, 학습할 입력 특징들을 설정한다. 그런 다음, 스케일링 과정을 통해 수치형 특성들의 스케일을 조정하여 모든 특성이 동등한 중요도를 갖도록 한다. 이는 모델 학습의 효율성을 높이는 데 큰 역할을 한다. 그리고, 모델이 새로운 데이터에 대한 일반화 능력을 갖추도록 하기 위해 데이터를 학습과 테스트 세트로 분할하는데, 이번 연구에서는 학습 세트는 전체 데이터의 80%, 테스트 세트는 20%로 분할했다. 이렇게 분할된 데이터 세트는 각각 모델을 학습시키고 모델의 일반화 성능을 평가하는 데 사용된다.

3.2 Binary relevance method

본 연구는 다중 레이블 분류 문제에 대해 BR 방법을 적용하여 접근함으로써, 각 레이블을 독립적인 이진 분류 문제로 변환하여 처리하는 방법론의 효율성과 한계를 탐구하였다. BR 방법론은 복잡한 다중 레이블 데이터를 각각의 레이블에 대해 단일 레이블 데이터로 분할함으로써, 각 레이블마다 독립적인 분류기를 구축하고 학습시키는 절차를 통해, 모델의 학습과 예측 과정을 단순화하는 장점을 제공한다[4][5]. 다만 BR 방법의 근본적인 한계로는 레이블 간 상호작용이나 상관관계를 고려하지 않음으로써 발생하는 정보 손실 문제가 지적되어 왔다.

이번 연구 차원에서 BR 적용 과정을 구체적으로 설명하면 다음과 같다. 먼저 전체 데이터 세트로부터 각 질환 만을 타겟으로 하는 9개의 질환 별 데이터 세트를 생성하고, 각 레이블은 해당 질환이 양성인 경우 1, 음성인 경우는 0으로 설정한다. 그리고 각 질환 마다 독립된 이진 분류기를 학습 시키는데 질환 별 최적 모델을 찾기 위해서 다음에 설명되는 다양한 알고리즘들을 학습 시킨다. 이를 통해서 한 건강 검진 데이터가 주어졌을 때 각 질환 분류기는 레이블에 속하는 지를 예측하고, 이렇게 9개 분류기의 예측 결과가 나오면 이를 모아서 최종 레이블 세트를 완성한다. 이러한 BR 모델의 최종 정확도에는 각 레이블에 대한 예측이 실제 레이블과 얼마나 일치하는지를 기준으로 다음에 설명되는 평가 지표를 사용한다.

3.3 분류 알고리즘

본 연구는 다양한 분류 알고리즘을 사용해서 모델을 학습하고 성능을 비교함으로서 이번 데이터에

맞는 최적의 모델을 찾고자 하였다. 여러 알고리즘 중 대표적인 로지스틱 회귀, 서포트 벡터 머신, 랜덤 포레스트, 그래디언트 부스팅 등 다섯 가지 분류 알고리즘을 사용하였다.

로지스틱 회귀(LR, Linear Regression)은 선형 모델에 기반한 방법으로 각 그룹별 확률 추정치를 계산하여 임계치에 따라 데이터를 그룹으로 분류한다[10]. 이는 주로 이진 분류 문제에 적용되며, 결과는 시그모이드 함수를 통해 0과 1 사이의 값으로 출력된다. LR에 대한 수식은 식 (1)과 같다.

$$\min_\theta -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))] \quad (1)$$

여기서 x와 y는 입력값이며, m은 데이터의 수, h는 비용 함수이다.

서포트 벡터 머신(SVM, Support Vector Machine)은 초평면(Hyperplane)을 기준으로 동일한 특성을 갖는 데이터를 분할하여 분류 분석을 수행한다[11]. 이 알고리즘은 마진(Margin) 최대화를 통해 최적의 결정 경계를 찾으며, 비선형 데이터에도 커널 트릭을 적용하여 사용될 수 있습니다.

랜덤 포레스트(RF, Random Forest)는 다수의 결정 트리를 기반으로 하는 앙상블 학습 방법이다[12]. 이 방법은 무작위로 선택된 서브셋에서 결정 트리를 각각 독립적으로 훈련시키고, 최종 예측은 이들 트리의 결과를 통합하여 결정한다. RF는 무작위성을 통해 각 트리가 서로 다른 데이터의 특성을 학습함으로써 과적합을 방지하고, 노이즈에 강한 내성을 갖는다. 또한, 부트스트랩 집계(배깅) 방식을 사용하여 서로 다른 훈련 데이터 세트으로 훈련된 각기 다른 트리들을 결합함으로써 모델의 안정성과 정확도를 향상 시킨다.

그래디언트 부스팅(GB, Gradient Boosting)은 앙상블 방법 중 하나로, 기울기 강하를 사용하여 최적의 파라미터를 찾는 방법이다[13]. 이 알고리즘은 여러 모델의 예측을 결합하여 성능을 향상시키며, 강력한 성능을 보이지만, 파라미터 설정과 학습 시간 면에서는 단점을 가진다. 특히, 학습률, 부스팅 스테이지 수, 트리 깊이, 나뭇잎 수 등의 파라미터는 다른 알고리즘보다 민감하다.

3.4 성능 평가 지표

평가 지표들을 활용하여 각 레이블별 모델 성능의 정밀한 분석을 수행하고, 이를 바탕으로 모델의 개선 가능성을 탐색한다. 이러한 지표들은 분류 모델의 다양한 성능 측면들을 폭넓게 평가하는 데 중요한 역할을 하며, 실제 환경에서의 응용을 고려할 때 특히 중요한 정보를 제공한다[14]. 이 과정은 데이터 분석과 머신 러닝 모델의 성능을 체계적으로 이해하고 평가하는 데 있어 필수적이며, 모델의 향후 발전 방향에 대한 중요한 가이드라인을 제공한다.

이진 분류 모델의 성능을 이해하고 정량적으로 평가하기 위해 혼동 행렬(Confusion matrix)은 중요하다. 혼동 행렬은 단순히 정확도(Accuracy)만으로 파악하기 어려운 모델의 성능에 대한 정보를 제공하며, 특히 클래스 불균형이 있는 경우에 유용하다. 또한, 이를 통해 모델이 특정 클래스를 과도하게 예측하는 경향이 있는지 확인할 수 있다. 혼동 행렬은 다양한 성능 지표를 계산할 수 있으며, 정밀도, 재현율, F1-score도 이에 해당된다. 표 2에 나와 있듯이, 혼동 행렬은 모델의 예측 성능을 시각적으로 이해하기 위한 중요한 도구로, 실제 결과와 예측 결과를 비교하여 제공한다. 이진 분류에 대한 혼동 행렬은 참 양성(TP, True Positive), 거짓 양성(FP, False Positive), 거짓 음성(FN, False Negative), 그리고 참 음성(TN, True Negative)으로 구성된다.

표 2. 이진 분류 혼동 행렬
Table 2. Confusion matrix for binary classification

		Predicted:	
		Negative(N)	Positive(P)
Actual:	Negative (N)	True Negative (TN)	False Positive (FP)
	Positive (P)	False Negative (FN)	True Positive (TP)

모델의 성능을 나타내는 데에는 보통 정확도(Accuracy)가 사용되지만, 정확도는 불균형 데이터 세트에서 다수 클래스의 예측에 대해서 치우침이 발생할 수 있으므로 추가적인 지표들을 사용한다.

정밀도는 모델이 양성으로 예측한 경우 중 실제로 양성인 비율을 나타내는 지표로, 다음과 같이 계산된다.

$$PR = \frac{TP}{TP+FP} \quad (2)$$

재현율은 실제 양성 케이스 중 모델이 올바르게 양성으로 예측한 비율을 나타내는 지표로, 다음과 같이 계산된다.

$$RC = \frac{TP}{TP+FN} \quad (3)$$

F1 점수는 정밀도와 재현율의 조화 평균을 나타내는 지표로, 두 지표의 균형을 평가하는 데 사용되며, 다음과 같이 계산된다.

$$F_1 = 2 \times \frac{PR \times RC}{PR+RC} \quad (4)$$

F1은 BR 다중 레이블 분류에서 각 레이블에 대한 예측 정확성을 고려할 수 있으며, 클래스 간 불균형이 존재하는 경우에도 신뢰할 수 있는 성능 측정을 제공한다.

예측에서 음성(TN)을 잘 맞추는 지도 중요한데 평가하는 지표로는 주로 '특이도(Specificity)'가 사용된다. 특이도는 실제로 음성인 경우 중에서 모델이 음성으로 정확하게 예측한 비율을 나타내면 다음과 같이 계산할 수 있다.

$$SP = \frac{TN}{TN+FP} \quad (5)$$

Ⅳ. 결과 및 토론

4.1 모델 선택

본 연구에서의 분석은 Python 계열의 Scikit-learn [15] 라이브러리를 활용하여 III 장에서 설명한 로지스틱 회귀(LR), 서포트 벡터 머신(SVM), 랜덤 포레스트(RF), 그리고 그래디언트 부스팅(GB) 알고리즘을 학습 세트로 훈련시키는 작업에서 시작되었다. 이 과정에서 GridSearchCV를 활용한 하이퍼파라미터 최적화는 모델의 성능을 미세 조정하는 데 핵심적인 역할을 수행하였다. 'n_estimators', 'max_depth'등의 주요 매개변수를 최적화하여, 다양한 하이퍼파라미터 조합을 실험한 결과, 성능이 가장 우수한 조합을 도출해내는 모델 선택 절차가 이루어졌다. K-fold 교차 검증 방식을 채택하여 얻은 최적의 하이퍼파라미터 조합은 모델의 선택에 직접 적용되었으며, 이 방법을 통해 모델이 실제 환경에서의 성능을 예측하는 데에 대한 신뢰성을 높였다.

4.2 학습 성능 결과

III 장에 설명한 그리드 서치를 통한 BR 다중 레이블 분류 방식으로 학습한 결과가 다음에 설명된다. 표 3에서는 앞서 언급하였던 다양한 알고리즘별로 하이퍼파라미터 범위가 제시되어 있다.

표 3. 그리드 서치에서 알고리즘 별 하이퍼파라미터 범위
Table 3. Algorithm-specific hyperparameters in grid search

Algorithm	Range of hyperparameters
LR	C: [0.01, 0.1, 1], solver: ['newton-cg', 'lbfgs', 'liblinear']
SVM	C: [0.1, 1, 10], kernel': ['poly', 'rbf', 'sigmoid'], gamma': ['scale', 'auto']
RF	n_estimators: [100, 150, 200], min_samples_leaf: [1, 2, 3]
GB	learning_rate=[0.001, 0.01, 0.1], min_samples_split=[2, 3, 4], n_estimators=[50, 100, 150]

모든 질환에 대한 최적화된 파라미터 설정을 적용한 모델들의 각 지표별 성능은 테스트 데이터 세트를 기준으로 평가되었으며, 해당 결과는 표 4에 상세하게 제시되어 있다. 비만의 경우는 랜덤포레스트(RF)가 최고의 모델이며, 양성 예측율은 F1 기준 0.8519로 높은 수준이며 음성 예측율은 SP 기준 0.9890으로 매우 높은 수준이다. 고혈압의 경우는 그래디언트 부스팅(GB)가 최고의 모델이며, 양성 예측율은 F1 기준으로 0.6250으로 그리 높지 않은 수준이며 음성 예측율은 SP 기준 0.9821으로 매우 높은 편에 속한다.

당뇨의 경우는 그래디언트 부스팅(GB)가 최고의

모델이며, 양성 예측율은 F1 기준으로 0.0000으로 전혀 맞추지 못 하는 수준이며 음성 예측율은 SP 기준 0.9744으로 양성 예측율과 반대의 결과를 나타낸다. 근골격 불균형의 경우는 그래디언트 부스팅(GB)가 최고의 모델이며, 양성 예측율은 F1 기준으로 0.8421으로 잘 맞추는 수준이며 음성 예측율은 SP 기준 0.0000으로 음성 데이터에 대해 아예 못 맞추는 것을 볼 수 있다. 이 계산 방식은 다중 레이블 분류 문제에서 각 레이블별 예측의 정확도와 재현율을 종합적으로 고려함으로써 전체적인 모델 성능의 정밀한 평가를 가능하게 한다. 이 결과는 데이터셋의 특성, 분석 방법, 하이퍼파라미터 조정 및 데이터 전처리 방법에 따라 성능이 달라질 수 있음을 보여주며, 각 알고리즘의 특성에 맞는 적절한 모델 선택의 중요성을 강조한다.

본 연구에서는 학습 곡선을 통해 모델 학습 시 성능 개선 과정을 종합적으로 분석하였다. F1 점수 기준의 학습 곡선을 통해 훈련 데이터 양에 따른 모델의 학습 진행 상황과 성능을 평가하며, 이러한 과정을 통해 모델의 신뢰성을 높였다. 그림 1을 살펴보면, 비만에 대해 최고 모델인 RF의 학습 곡선에서 데이터 양이 증가함에 따라 validation score의 F1-score가 상승하고 동시에 수렴함으로써 모델의 예측 성능이 더 나아지는 것을 보여 주고 있다. 이는 모델이 더 많은 데이터로부터 학습하여 일반화 능력을 향상시키고 있음을 보여 준다.

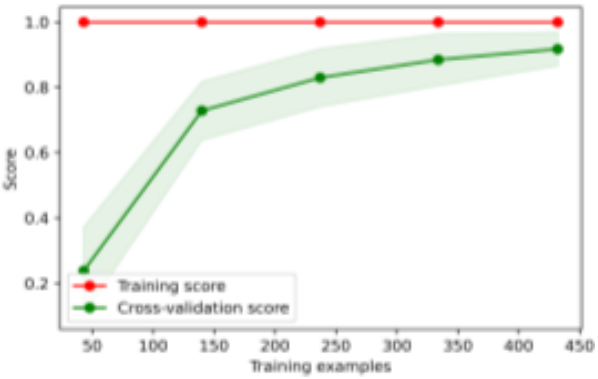

그림 1. 비만에 대한 RF의 학습 곡선
Fig. 1. Learning curve of obesity in RF

표 4. 테스트 데이터 세트를 사용한 질환 별 최고 이진 분류 알고리즘 및 성능
Table 4. Performance results of multi-label classification evaluated with the test data set

Illness	Best algorithm	Hyperparameters	PR	RC	F1	SP
Obesity	RF	'min_samples_leaf': 1, 'n_estimators': 150	0.9583	0.7667	0.8519	0.9890
Liverdisease	GB	'learning_rate': 0.1, 'min_samples_split': 2, 'n_estimators': 100	0.8824	0.6818	0.7692	0.9798
Hypertension	GB	'learning_rate': 0.1, 'min_samples_split': 2, 'n_estimators': 100	0.7143	0.5556	0.6250	0.9821
Diabetes	GB	'learning_rate': 0.1, 'min_samples_split': 2, 'n_estimators': 50	0.0000	0.0000	0.0000	0.9744
Kidneydisease	GB	'learning_rate': 0.01, 'min_samples_split': 2, 'n_estimators': 100	1.0000	0.5833	0.7368	1.0000
Dyslipidemia	GB	'learning_rate': 0.1, 'min_samples_split': 3, 'n_estimators': 50	0.0000	0.0000	0.0000	0.9744
Cardiovascular disease	LR	'C': 1, 'solver': 'newton-cg'	0.0000	0.0000	0.0000	0.9832
Musculoskeletal imbalance	GB	'learning_rate': 0.001, 'min_samples_split': 2, 'n_estimators': 50	0.7273	1.0000	0.8421	0.0000
Stress	SVM	'C': 10, 'gamma': 'scale', 'kernel': 'sigmoid'	0.3438	0.2973	0.3188	0.7500

본 연구에서는 레이블 별 이진 분류 모델의 양성 예측 성능을 평가하기 위해 ROC(Receiver Operating Characteristic) 커브와 AUC(Area Under the ROC Curve) 값을 활용하였다[16]. ROC 커브는 모델이 다양한 분류 임계값을 적용했을 때의 거짓 양성 비율(FPR, False Positive Rate) 대 진짜 양성 비율(TPR, True Positive Rate)의 변화를 시각적으로 나타내며, 모델이 양성 클래스를 얼마나 잘 분류하는지에 대한 중요한 정보를 제공한다. 이와 함께, AUC 값은 모델의 성능을 하나의 숫자로 요약하여, 모델이 실제 양성 클래스를 얼마나 정확하게 예측하는지에 대한 정량적인 평가를 가능하게 한다. 그림 2는 질환 별 ROC 커브를 보여 주는데, 비만과 간 질환에 대해서는 96% 이상의 매우 정확한 양성 예측을 보이고 있고, 고혈압, 신장 질환 및 이상지질혈증에 대해서는 70~80% 수준의 높은 정확도를 보이고 있고, 당뇨, 순환기 질환, 근골격 불균형 및 스트레스에 대해서는 70% 이하의 예측 성능을 보이고 있다. 이러한 분석을 통해서 향후 초기 모델을 어떤 방향으로 향상시 켜야 할지 방향성을 시사한다.

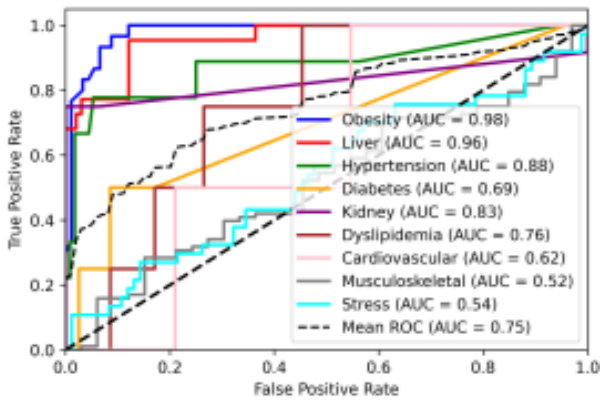

그림 2. ROC AUC 커브
Fig. 2. ROC AUC curve

질병 별 혼동 행렬은 지면 제한 상 대표 질환 세 개에 대해서 나타냈다. 표 5는 비만 질환에 대한 혼동 행렬이며, test set로 분리한 결과, 음성 데이터는 91개가 있고 양성 데이터는 30개가 있으며, 각각 1개와 7개의 데이터를 잘못 예측한 것을 확인할 수 있다. 여기서 볼 수 있듯이, 해당 질환의 음성과 양성 데이터는 크게 불균형하다는 것을 알 수 있다. 본 질환은 양성 데이터가 음성 데이터에 비해 정확도가 떨어지지만, 전반적으로 높은 예측률을 지니는 것을 알 수 있다.

표 5. 비만에 대한 혼동행렬
Table 5. Feature importance for obesity

		Predicted:	
		N	P
Actual:	N	90	1
	P	7	23

표 6은 간 질환에 대한 혼동 행렬이며, test set로 분리한 결과, 음성 데이터는 99개가 있고 양성 데이터는 22개가 있으며, 각각 2개와 7개의 데이터를 잘못 예측한 것을 확인할 수 있다. 여기서 볼 수 있듯이, 해당 질환의 음성과 양성 데이터는 크게 불균형하다는 것을 알 수 있다. 본 질환은 음성 데이터에 대해 대부분 맞췄지만, 양성 데이터가 극히 부족함으로써 절반 정도의 데이터만 맞춘 것으로 볼 수 있다.

표 6. 간 질환에 대한 혼동행렬
Table 6. Feature importance for liver

		Predicted:	
		N	P
Actual:	N	97	2
	P	7	15

표 7은 고혈압에 대한 혼동 행렬이며, test set로 분리한 결과, 음성 데이터는 112개가 있고 양성 데이터는 9개가 있으며, 각각 2개와 4개의 데이터를 잘못 예측한 것을 확인할 수 있다. 여기서 볼 수 있듯이, 해당 질환의 음성과 양성 데이터는 크게 불균형하다는 것을 알 수 있다. 본 질환은 앞서 언급되었던 질환 들과 같이 음성이 양성 데이터보다 많아 음성에 대한 정보들을 잘 맞추지만 음성보다 데이터가 10배 이상 적은 양성 데이터에 대해 잘 못 맞춘 것으로 확인된다.

표 7. 고혈압에 대한 혼동행렬
Table 7. Feature importance for hypertension

		Predicted:	
		N	P
Actual:	N	110	2
	P	4	5

4.3 토론

이번 연구는 국내에서 건설-제조 현장 작업자들을 대상으로 한 대규모 건강 검진 데이터를 사용한 최초 다중 레이블 분석 시도로서 의미가 있다. 본 연구에서는 의료 현장의 필요성에 의해서 대표적인 9개 성인 질병 레이블과 82개 건강 측정 항목을 모두 사용해서 다중 레이블 분류를 시도해 보았다. 종합적으로 보았을 때 너무 낮은 양성 예측률을 보이는 질환 경우가 많아서 실제 임상 적용에 충분한 수준은 아닌 것으로 판단된다. 다만 특히 의료 분야에서 음성을 정확하게 예측하는 성능은 신뢰성이나 비용 절감 면에서 아주 중요한데, 이번 모델의 음성 예측 성능이 매우 높은 편이어서 초기 분석 모델로서 유의미한 결과라고 판단된다. 향후 임상에 적용할 정도의 정밀한 BR 모델을 만들기 위해서는 도메인 전문가인 의료진들과 협업을 통해서 질환 별로 이진 분류의 최적화 모델을 각각 찾아서 너무 성능이 낮은 레이블은 배제하거나 다시 레이블링을 하고, 측정 항목 특징 중요도를 계산해서 불필요한 특징을 배제하고 모델을 학습한다면 예측 성능을 보다 높일 수 있음은 물론 안정된 모델을 얻을 수 있을 것이다.

또한 수집된 데이터는 자연스럽게 레이블 간 불균형 상태를 보이고 있는데, 이번에는 초기 연구라 데이터 왜곡이나 정보 손실을 최소화하기 위해서 학습 데이터를 인위적으로 균형 있게 만들지는 않았고, 대신 불균형 문제 대처에 효과적인 앙상블 학습을 포함했고 대부분의 질환에서 앙상블 분류기가 높은 성능을 보였다. 향후 불균형 데이터 문제에 좀 더 적극적으로 대처하기 위해서 추가적인 데이터의 수집은 물론 랜덤 리샘플링이나 합성 소수 오버샘플링 기법(SMOTE, Synthetic Minority Oversampling Technique) 등의 적용도 가능할 것이다. 다만 오버샘플링 시 데이터 왜곡이나 다운샘플링 시 정보 손실을 최소화 하기 위해서는 다양한 방법을 시도하고 서로 비교해서 주의 깊게 적용해야 한다[17].

Ⅴ. 결론 및 향후 과제

본 연구의 기여는 다음과 같다. 첫째, 최초로 건설-제조 현장 작업자를 대상으로 장기간 대규모로 수집된 정밀 건강 검진 데이터를 사용해서 복합성 질환 예측을 위한 다중 레이블 데이터 분석을 시도했다. 둘째, 다양한 분류 알고리즘들을 그리드 서치 기반의 하이퍼파라미터 튜닝을 통해서 학습시키면서 최적의 모델을 선택하는 방법을 작업자의 복합 질환과 연관된 건강 예측에 응용하고 체계화했다. 셋째, 질병들 중 불균형이 있는 경우에 모델 선택에 있어서 적절한 성능 지표를 기반으로 진단 정확도를 향상시켜서 드문 질병을 포함하여 다양한 건강 상태를 효과적으로 구별할 수 있게 했다.

본 연구의 한계점은 다음과 같다. 첫째, 이진 관련성 접근법은 레이블 간의 상관 관계를 무시할 수 있기 때문에 그러한 관계의 고려가 중요할 경우 classifier chains, label powerset, deep learning 등을 이용해서 좀 더 정밀한 모델링이 필요할 수 있다. 둘째, 데이터 불균형 문제나 과적합의 위험을 줄이기 위해서는 좀 더 많은 데이터가 필요할 수 있는데, 실제 데이터를 추가하는 것은 상당한 기간과 비용을 초래할 수 있어서 쉽지 않은 일이다. 기존 데이터 세트를 이용해서 의료 데이터의 민감성과 정확성을 왜곡하지 않도록 주의하면서 GANs (Generative Adversarial Networks) 같은 생성형 모델을 이용한 데이터 증강을 시도해 볼 수 있다.

개인 의무 기록을 기반으로 인공지능 기술을 통한 질병 예측 모델링이 가능해 지면서 의료 기술과 데이터 분석을 통한 인공지능 기술의 상호 융합을 통해 특수 산업 현장의 위험 요인을 발견하는 것은 혁신적인 헬스케어 산업의 부가가치 창출에 대한 기대감을 높여주고 있다. 특히, 근로자의 위험 건강 요인의 발견을 통한 산업 재해의 예방과 업무 생산성 향상에 대한 기대는 4차산업혁명의 기술이 기존 산업과 융합하여 얻을 수 있는 긍정적 신호라 할 수 있다.

이에 본 연구에서는 건설 · 제조 근로자의 건강 위험 관찰을 위한 머신러닝 기법을 제안하였으며, 도출된 건강 위험 요인의 특징 우선순위 발견을 통해 건강 관리 서비스로의 가치도 발견하였다. 본 연구의 결과를 통해 국내 유관 산업 및 헬스케어 분야에 종사하는 스타트업, 병원, 기업, 연구소 등이 특수 산업 현장의 위험 요인을 발견하는 인프라가 확대될 것이라 기대한다.

특히, 인공지능 기술을 통해 산업 내 재배 비용 절감 · 업무 생산성 향상 · 효율적 기업 운영 등 전반적인 산업 내 성장 및 효율성이 증가될 것이며, 이를 위해 개인 의무 기록이 안전하게 관리 · 활용되고, 산업 내 맞춤형 근로자 지원 프로그램으로 성장하길 기대한다.

References

[1] Y. E. Min, S. J. Cho, J. H. Jeong, Y. W. Sung, and Y. Shin, "Predicting Model for Occupational Disease using Complex Analysis in Construction Site", The Korean Society of Science & Art, Vol. 36. pp. 109-120, Dec. 2018. http://dx.doi.org/10.17548/ksaf.2018.12.30.109.

[2] Y.-M. Lim, Y.-O. Kwon, and W.-S. Kim, "Design and Application of Disease Identification for u-Health Environment", The Journal of Korean Institute of Information Technology, Vol. 8, No. 2, pp. 101-107, Feb. 2010.

[3] S. B. Lee, J. Song, and A. Park, "A Trend of Artificial Intelligence in the Healthcare", The Journal of the Korea Contents Association, Vol. 20, No. 5, pp. 448-456, May 2020. https://doi.org/10.5392/JKCA.2020.20.05.448.

[4] R. Li, W. Liu, Y. Lin, H. Zhao, and C. Zhang, "An Ensemble Multilabel Classification for Disease Risk Prediction", Journal of Healthcare Engineering, Vol. 2017, No. 2, pp. 1-10, Jun. 2017. http://dx.doi.org/10.1155/2017/8051673.

[5] A. N. Tarekegn, M. Giacobini, and K.Michalak, "A review of methods for imbalanced multi-label classification", Pattern Recognition, Vol. 118, pp. 107965, Oct. 2021. https://doi.org/10.1016/j.patcog.2021.107965.

[6] M. L. Zhang and L. Wu, "Lift: Multi-label learning with label-specific features", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 37, No. 1, pp. 107-120, Jan. 2015. https://doi.org/10.1109/TPAMI.2014.2339815.

[7] J. Huang, G. Li, Q. Huang, and X. Wu, "Learning Label Specific Features for Multi-Label Classification", IEEE International Conference on Data Mining, Atlantic City, NJ, USA, pp. 181-190, Nov. 2015. https://doi.org/10.1109/ICDM.2015.67.

[8] E. Alvares-Cherman, J. Metz, and M. C.Monard, "Incorporating label dependency into the binary relevance framework for multi-label classification", Expert Systems with Applications, Vol. 39, No. 2, pp. 1647-1655, Feb. 2012. https://doi.org/10.1016/j.eswa.2011.06.056.

[9] S.-H. Byun, "A study on utilization of healthcare application based on disease prediction algorithm -Focusing on prediction of national 5 major cancers and 12 major diseases", The e-Business Studies, Vol. 24, No. 5, pp. 119-137, Oct. 2023.

[10] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review", Journal of biomedical informatics, Vol. 35, No. 5, pp. 352-359, Oct. 2002. http://dx.doi.org/10.1016/s1532-0464(03)00034-0.

[11] M. Awad and R. Khanna, "Supportvector machines for classification", Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, pp. 39-66, Apr. 2015. https://doi.org/10.1007/978-1-4302-5990-9_3.

[12] L. Breiman, "Random forests", Machine learning, Vol. 45, pp. 5-32, Oct. 2001. https://doi.org/10.1023/A:1010933404324.

[13] J. H. Friedman, "Stochastic gradient boosting", Computational statistics & data analysis, Vol. 38, No. 4, pp. 367-378, Feb. 2002. https://doi.org/10.1016/S0167-9473(01)00065-2.

[14] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation", Journal of Machine Learning Technologies, Vol. 2, No. 1, pp. 37-63, Feb. 2011.

[15] F. Pedregosa, et al., "Scikit-learn: Machine learning in Python", Journal of Machine Learning

Research, Vol. 12, pp. 2825-2830, Nov. 2011.

[16] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms", Pattern recognition, Vol. 30, No. 7, pp. 1145-1159, Jul. 1997. https://doi.org/10.1016/S0031-3203(96)00142-2.

[17] L. Nanni, C. Fantozzi, and N. Lazzarini, "Coupling different methods for overcoming the class imbalance problem", Neurocomputing, Vol. 158, pp. 48-61, Jun. 2015. https://doi.org/10.1016/j.neucom.2015.01.068.

저자소개

등 덕 호 (Teh-Hao Teng)

2017년 9월 : 고려대학교 기계공학부(공학사)
2023년 9월 ~ 현재 : 성균관대학교 지능형로봇학과 석사 과정
관심분야 : 빅데이터 분석, 로보틱스

이 태 림 (Tae-Rim Lee)

2018년 8월 : 한국외국어대학교 (공학사)
2023년 2월 ~ 현재 : 성균관대학교 인공지능학과 석.박사 통합 과정
관심분야 : 머신러닝, 딥러닝, 강화학습

변 승 혁 (Seung Hyeok Byeon)

2018년 : 동국대학교 경제학박사
2019년 ~ 현재 : ㈜인피니티케어 R&D연구소 연구소장
2019년 ~ 2023년 : 서울특별시 건강전문가 자문위원(DB분석 및 건강관리 지표 체계 부문)
2023년 ~ 현재 : 동국대학교 글로벌무역학과 대우교수
관심분야 : 빅데이터, 개인의무기록, 마이데이터, 질병예측

최 문 택 (Mun-Taek Choi)

2000년 5월 : U. of Southern California 항공우주공학과 (공학박사)
2013년 3월 ~ 현재 : 성균관대학교 지능형로봇학과 교수
관심분야 : 로봇 지능 학습, 헬스케어 빅데이터 분석, 근골격 재활 분석, 외골격 로봇 제어