https://doi.org/10.4306/jknpa.2017.56.3.103

ORIGINAL ARTICLE

J Korean Neuropsychiatr Assoc
2017;56(3):103-110
Print ISSN 1015-4817
Online ISSN 2289-0963
www.jknpa.org

조현병 감별진단에 대한 머신 러닝 기법의 적용 : WAIS-IV의 진단 예측 역량

을지대학교 을지대학교병원 정신건강의학과,[1] 동국대학교 일산병원 정신건강의학과[2]
고은해[1] · 강희양[1] · 김용식[2] · 정성훈[1]

A Case Study of a Machine-Learning Approach in Differential Diagnosis of Schizophrenia : The Predictive Capacity of WAIS-IV

Eun Hae Ko, MD[1], Hi Yang Kang, PhD[1],
Yong Sik Kim, MD, PhD[2], and Seong Hoon Jeong, MD, PhD[1]

[1]Department of Psychiatry, Daejeon Eulji Medical Center, Eulji University, Daejeon, Korea
[2]Department of Psychiatry, Dongguk University Ilsan Hospital, Goyang, Korea

Objectives Machine learning (ML) encompasses a body of statistical approaches that can detect complex interaction patterns from multi-dimensional data. ML is gradually being adopted in medical science, for example, in treatment response prediction and diagnostic classification. Cognitive impairment is a prominent feature of schizophrenia, but is not routinely used in differential diagnosis. In this study, we investigated the predictive capacity of the Wechsler Adult Intelligence Scale IV (WAIS-IV) in differentiating schizophrenia from non-psychotic illnesses using the ML methodology. The purpose of this study was to illustrate the possibility of using ML as an aid in differential diagnosis.

Methods The WAIS-IV test data for 434 psychiatric patients were curated from archived medical records. Using the final diagnoses based on DSM-IV as the target and the WAIS-IV scores as predictor variables, predictive diagnostic models were built using 1) linear 2) non-linear/non-parametric ML algorithms. The accuracy obtained was compared to that of the baseline model built without the WAIS-IV information.

Results The performances of the various ML models were compared. The accuracy of the baseline model was 71.5%, but the best non-linear model showed an accuracy of 84.6%, which was significantly higher than that of non-informative random guessing (p=0.002). Overall, the models using the non-linear algorithms showed better accuracy than the linear ones.

Conclusion The high performance of the developed models demonstrated the predictive capacity of the WAIS-IV and justified the application of ML in psychiatric diagnosis. However, the practical application of ML models may need refinement and larger-scale data collection.

J Korean Neuropsychiatr Assoc 2017;56(3):103-110

KEY WORDS Machine learning · Schizophrenia · WAIS-IV · Neuropsychological function · Diagnostic support system.

Received April 6, 2017
Revised May 11, 2017
Accepted May 26, 2017

Address for correspondence
Seong Hoon Jeong, MD, PhD
Department of Psychiatry,
Daejeon Eulji Medical Center,
Eulji University,
95 Dunsanseo-ro, Seo-gu,
Daejeon 35233, Korea
Tel +82-42-611-3443
Fax +82-42-611-3443
E-mail AnselmJeong@gmail.com

서 론

머신 러닝(machine learning 혹은 기계 학습)이라는 개념은 인공지능의 한 분야로, 데이터에 내재하는 패턴을 컴퓨터가 학습할 수 있도록 하는 제반 알고리즘을 지칭한다.[1)] 머신 러닝은 데이터 마이닝과 개념상 유사하나, 후자가 변수 간의 새로운 연관 관계를 발견하는 데 중점을 둔다면, 전자는 결과를 예측하는 실용적 목표를 추구한다.[2)] 이러한 목표의 차이 때문에 두 기법은 접근 전략도 차이가 난다. 데이터 마이닝을 비롯한 기존 통계분석에서는 선험적 통계 모델을 미리 정해놓고 데이터가 이를 뒷받침하는지 확인하고자 한다. 이에 비해 머신 러닝은 어떠한 선험적 가설도 없이, 데이터를 잘 예측해주기만 한다면 어떤 모델도 수용한다.[3)]

의학 연구에서는 수많은 변수가 고려되어야 하며, 이들 간의

상호작용 역시 간과되어선 안 된다. 예를 들어 조현병은 원인 요인이 극히 다양하며, 이들 요인 간의 복잡다단한 상호작용이 병태생리를 결정지을 것으로 예상하고 있다.[4] 따라서 선험적 가설을 검증하기 위해 데이터를 이용하는 기존의 연구 방법은 쉽게 한계에 봉착하며, 변수들 사이의 패턴을 인식해 내는 학습 알고리즘이 중요성을 얻게 되었다.[5] 이러한 원인 때문에 머신 러닝은 조현병 연구 특히 신경영상학,[4] 뇌파 등 생체신호 분석,[6,7] 약물 반응이나 임상 경과의 예측 등에 활발히 사용되기 시작하였다.[8,9]

전통적으로 조현병 진단은 병력 청취와 임상 면담을 통해 이루어진다. 조현병의 생물학적 진단 표지자는 발견되지 않았으며, 확진을 위한 진단법 개발 또한 요원하다.[10,11] 한편 정신과 진료실에서 비교적 손쉽게 행해지는 심리학적 검사에는 대부분 웩슬러 성인 지능검사(Wechsler Adult Intelligent Scale, 이하 WAIS)가 포함되어 있으며, 2008년 WAIS-IV가 도입된 이후 좀 더 세분된 영역에서 인지 기능을 평가하는 것이 가능해졌다.[12] 인지 기능 손상이 조현병의 핵심 증상이며, 주된 병태생리인 신경 회로망 장애를 더욱 밀접하게 반영한다고 볼 때,[13,14] 인지 증상을 진단 과정에 포함하면 조현병 감별진단에 큰 도움이 되리라고 여겨진다. 따라서 조현병 진단 기준에 인지 기능 손상이 포함되어야 한다는 주장이 끊임없이 제기되고 있지만,[14,15] 아직 정신장애 진단 통계 편람 5판(Diagnostic and Statistical Manual of Mental Disorder, 이하 DSM-5)은 진단 기준에 인지 기능 손상을 포함하지 않고 있다.[16]

한편 진단적 목적으로 인지 기능을 사용하는 데는 몇 가지 어려움이 있다. 과거 연구에서 환자군과 정상 대조군 사이에는 1.5~2.0 표준편차 정도의 인지 기능 차이가 있는 것으로 나타났다.[17] 하지만 제반 영역에 골고루 기능 손상이 보여 조현병에 특정한 패턴을 확인하기 어려웠다.[18,19] 또한, 인지 장애는 양극성 장애나 주요 우울 장애의 관해 상태에서도 흔히 나타나기 때문에, 더더욱 감별진단에 이용되기 어려워졌다.[15,20] 따라서 단순히 개개 인지 기능을 따로따로 비교하는 일변량 통계분석은 감별진단을 내리는 데는 역부족일 것이다. 이상적으로는 관련 변수 간의 비선형 관계를 고려하여 복합적인 패턴을 찾는 알고리즘이 필요하다. 머신 러닝 알고리즘은 이러한 패턴을 찾는 방법들 중 하나이다.[21] 실제로 최근 머신 러닝 알고리즘이 대중화되면서 예측 모델링의 성능이 크게 개선되었으며 의학적 진단 목적으로도 상당한 성공을 거두고 있다.[21,22]

웩슬러 성인 지능검사는 광범위하게 사용될 뿐 아니라[23] 동시에 가장 활발히 연구된 도구 중 하나이다.[14,24,25] 조현병 환자는 정상 대조군보다 유의하게 지능 지수가 낮았으며,[25] 고전적 메타분석에서도 WAIS 검사는 조현병과 정상 대조군을 가장 잘 구분하는 도구 중 하나로 나타났다.[26] 따라서 WAIS 자료를 통하여 조현병 감별진단을 도울 가능성을 타진하는 것은 의미 있는 시도라 하겠다.

본 연구에서는 WAIS-IV 자료를 바탕으로, 정신병적 증상을 동반하지 않은 기타 질환 환자로부터 조현병 환자 감별진단을 도울 수 있는 예측 모델을 세워보았다. 머신 러닝을 이용한 방법론의 유용성을 입증하기 위해 본 연구에서는 3가지 예측 모델을 구축하여 서로 비교하였다 : 1) 인구 통계학적 변수만을 포함하는 기저 모델, 2) 선형 통계적 접근법을 사용하는 모델, 3) 비선형/비모수적 머신 러닝 알고리즘을 사용하는 모델. 이들 세 가지 종류 모델의 정확도를 서로 비교하기 위해, 학습에 사용되지 않은 테스트 자료를 이용하여 실제 임상과 유사한 상황에서 어느 정도의 정확성을 기대할 수 있는지 가늠해보았다. 만약 고전적 통계기법으로는 해결되지 않던 감별진단의 문제에 머신 러닝을 적용함으로써 해결의 실마리를 찾을 수 있다면, 머신 러닝 적용의 가능성 및 유용성을 입증하게 될 것이다.

방 법

자 료

본 연구는 일 대학병원 정신건강의학과에 입원 혹은 외래를 통해 내원한 환자의 의무 기록을 검토한 후향적 의무기록 조사 연구이다. 자료 수집 절차와 전반적인 연구 방법에 대해서는 기관윤리위원회의 검토를 받았다(승인번호 EMC 2017-05-001). 웩슬러 성인 지능검사가 도입된 이후 이를 시행한 모든 환자의 자료를 진단에 관계 없이 수집하였다. 이 중 검사 당시 나이가 16세 미만인 경우, 정신 지체, 기질적 뇌 질환, 인격장애 및 치매로 진단된 환자의 자료는 제외하였다. 군 복무 적합이나 산재 판정을 위해 검사를 받은 경우도 제외하였다. 수집된 대상자에 대하여 저자 중 일인인 정신건강의학 전문의가 의무기록을 재검토하였다. 검사를 행한 이후의 치료 경과를 토대로 하여, 정신 장애의 진단 및 통계 편람 제 4판(DSM-IV)에 따라 진단을 확인하였다. 진단에 따라 정신증(psychosis) 및 비정신증(non-psychosis)으로 크게 나누었고, 정신증 환자 중에서는 조현병으로 확진된 경우만을 남기고 모두 후속 연구에서 제외하였다. 이때 제외된 진단에는 조현정동장애, 조현양상장애, 단기 정신병적 장애, 기타 정신병적 장애가 포함되었다.

최종 분석에는 남성 252명과 여성 182명, 총 434명의 자료가 포함되었다. 이중 조현병으로 진단된 환자는 114명이었으며, 비정신증으로 진단된 환자는 320명이었다. 비정신증군의 진단은 표 1과 같다. 일부 기분 장애 환자에서 일시적으

Table 1. The final diagnoses of included subjects based on DSM-IV

Schizophrenia		Non-psychotic illness	
Diagnosis	n (%)	Diagnosis	n (%)
Schizophrenia	114 (100.0)	Adjustment disorder	45 (14.1)
		Anxiety disorder	54 (16.9)
		Bipolar disorder	18 (5.6)
		Depressive disorder	141 (44.1)
		Obsessive compulsive disorder	15 (4.7)
		Panic disorder	16 (5.0)
		Acute stress disorder/posttraumatic stress disorder	31 (9.7)
Total	114 (100.0)	Total	320 (100.0)

DSM : Diagnostic and Statistical Manual of Mental Disorder

로 정신병적 증상을 나타내는 경우가 있었지만 조현정동장애로 진단이 내려진 경우가 아니라면 비정신증 범주로 분류하였다.

평가 척도

웩슬러 성인 지능검사, 4판(Wechsler Adult Intelligence Scale, WAIS-IV)

웩슬러 성인 지능검사는 2008년에 미국에서 개정되었으며, 2012년에 국내에서 번안, 표준화되었다.[27] 검사대상 연령은 16~69세까지의 성인으로, 열 개의 소검사를 통해 얻어진 점수를 토대로 네 개의 요인, 즉 언어이해, 지각추론, 작업기억, 처리속도를 계산하며, 전반적인 지적 능력을 보여주는 전체 지능 지수를 산출한다. 각각의 소검사 점수는 평균 10, 표준편차 3인 표준점수로 보고되며, 네 요인 및 전체 지능 지수는 평균 100, 표준편차 15인 조합점수로 보고된다. 본 연구에서는 개개 소검사 점수는 물론, 네 개의 요인 점수, 전체 지능 지수를 입력 자료로 하여 예측 모델을 구축하였다.

통계 분석

조현병군과 비정신증군 간의 WAIS-IV 결과 비교

웩슬러 성인 지능검사를 구성하는 각각의 변수가 조현병군과 비정신증군 사이에 실제로 차이를 드러내는지 스튜던트 t 검정을 이용하여 확인하였다. 동시에 두 군 간의 차이를 가장 잘 드러내는 변수를 확인하기 위하여 Cohen's d를 이용한 효과 크기(effect size)를 구하였다.

기저 모델(Baseline model) 구축

기저 모델은 오로지 성별과 나이가 주어졌을 때 비정신증으로부터 조현병을 얼마나 감별할 수 있는지를 보여준다. 이는 아무런 정보가 없을 때의 예측 정확도를 나타내며, 다른 모델에 대한 비교 기준으로 삼았다. 기저 모델은 이항 로지스틱 회귀 모델(binary logistic regression, 이하 LR)을 이용하여 구축하였다.

머신 러닝 알고리즘

머신 러닝 알고리즘 중에서 가장 널리 사용되는 4개의 선형 알고리즘과 3개의 비선형(nonlinear)/비모수적(nonparametric) 알고리즘을 적용하였다. 선형 알고리즘으로는 1) 이항 로지스틱 회귀(LR), 2) 벌점 이항 로지스틱 회귀(penalized binary logistic regression, 이하 PLR), 3) 선형 서포트 벡터 머신(linear support vector machine), 4) 선형 판별 분석(linear discriminant analysis)을 이용하였으며, 비선형/비모수적 알고리즘으로는 1) 방사 기저함수 커널 서포트 벡터 머신(radial basis function kernel support vector machine, 이하 RBF-SVM), 2) 랜덤 포레스트(random forest), 3) 그래디언트 부스팅(gradient boosting, 이하 GBM)을 사용하였다.

모델 구축 과정

전체 자료를 모델 구축에 사용될 훈련 데이터 세트(training dataset)와 추후에 예측 정확도를 평가하기 위해 사용될 테스트 데이터 세트(test dataset)를 나누었다. 두 데이터 세트에는 동일한 비율로 두 진단군(조현병, 비정신증)이 분포될 수 있도록 조절하였다. 각각의 알고리즘에 해당하는 모델을 만들고 목표값과의 오차를 최소화하도록 학습시켰다. 각각의 알고리즘에는 학습의 효율을 결정하는 초매개변수(hyperparameter)를 정하기 위해 반복 5배 교차 검증(repeated cross-validation)을 사용하였다.

오차가 최소화된 상태를 예측 모델로 삼았다. 얻어진 모델에 앞에서 마련해놓은 테스트 데이터 세트를 입력하여 예측값을 산출한 후 정답과 비교하여 예측 정확도를 추정하였다. 얻어진 값은 실제 임상에서 모델이 제공할 수 있는 정확도를 간접적으로 반영한다.

오차를 평가하는 척도

예측 정확도는 퍼센트 일치도와 함께 카파 통계량(Cohen's kappa)을 이용하였다. 퍼센트 일치도는 가장 기본적인 지표이지만, 본 연구와 같이 표본 내 두 군의 분포가 불균형 상태에 있을 때는 결과를 왜곡할 수 있다. 이러한 불균형 상태를 보정한 것이 카파 통계량이며, 일반적으로 0.4 이상이면 중등도의 일치도를 보이는 것으로 본다.[28]

퍼센트 일치도의 통계적 유의성을 계산하기 위해, 전혀 정보가 없을 때의 일치도(no-information rate)에 비해 모델에서 얻어진 일치도가 큰지를 단측 검정으로 평가하였다.

사용된 소프트웨어

모든 통계 분석은 'R' 소프트웨어(버전 3.2.4, Vienna, Austria)를 사용하였다.[29] 머신 러닝 모델의 구축에는 R 패키지 'caret'를 사용하였다.[30] 'R' 소프트웨어는 범용 통계 분석과 그래프 작성, 정보 시각화를 위하여 사용되는 공개 소프트웨어로, 프로그래밍이 가능하고 기능의 확장성이 뛰어나 SPSS나 SAS를 대체할 수 있는 범용 통계 소프트웨어로 각광받고 있다. 'R' 소프트웨어는 다양한 패키지에 의해 기능이 확장되며, 머신 러닝의 기능을 구현해주는 패키지만 해도 40개 이상에 달하며 이 수는 매년 증가하고 있다. 그중 본 연구에 사용된 'caret' 패키지는 다양한 머신 러닝 관련 기능을 단일한 인터페이스를 통해 사용할 수 있도록 함으로써 머신 러닝 모델의 구축, 결과 해석 등을 간편하게 구현할 수 있도록 도와준다. 한편 'caret' 패키지는 다른 패키지를 이용할 수 있게 돕는 역할만 담당할 뿐으로, 개개 알고리즘을 구현하는 패키지 자체는 전세계 수많은 통계학자들이 개발하여 피어 리뷰(peer review)를 거쳐 무료로 공개된 것이다.

결 과

인구학적 특성

조현병군과 비정신증군의 인구학적 자료를 표 2에 요약하였다. 성 비율에는 유의한 차이가 없었으며[$\chi^2(1)=1.58$, p=0.208], 연령 및 교육 기간에서도 의미 있는 차이는 발견되지 않았다. 조현병군의 평균 연령은 31.1±11.6세였고 비정신증군의 평균은 32.9±14.6세였다[T(432)=1.23, p=0.218]. 교육 기간의 평균값은 조현병군이 13.4±3.8년, 비정신증군이 13.1±4.6년이었다[T(432)=0.539, p=0.590].

웩슬러 성인 지능검사 결과

웩슬러 성인 지능검사 결과를 표 3에 요약하였다. 웩슬러 성인 지능검사의 모든 소검사 및 네 개 요인, 그리고 전체 지능 지수까지 모두 조현병군이 유의하게 점수가 낮았다. 전체 지능 지수의 평균값은 조현병군 77.8±17.2, 비정신증군 91.5±17.1였다[T(432)=7.32, p<0.001]. Cohen's d로 살펴본 효과 크기는 0.290에서 1.003의 범위였으며, 효과 크기가 가장 큰 것은 처리 속도(d=1.003), 가장 작은 것은 상식(d=0.290)이었다.

기저 예측 모델의 예측 정확도

성별과 나이를 설명 변수로 한 이항 로지스틱 회귀모델에서 얻어진 퍼센트 일치도는 71.5%, 카파 통계량은 0.141로 무작위로 추측하는 경우와 유의한 차이를 보이지 않았다(p=0.76).

선형 모델의 예측 정확도

선형 알고리즘을 사용하였을 때 얻어진 퍼센트 일치도는 76.2~78.5% 정도였다(표 4). 이 중 어느 것도 무작위로 추측했을 때와 비교해서 통계적으로 유의한 차이를 보이지 않았다. 카파 통계량 역시 0.311~0.397 정도로 중등도 일치를 표시하는 기준인 0.4에 미치지 못하였다.

비선형/비모수 모델의 예측 정확도

비선형/비모수 알고리즘에서 얻어진 퍼센트 일치도는 최소 81.5%(GBM)에서 최대 84.6%(RBF-SVM)였다. 이들은 모두 무작위로 추측했을 때와 비교하여 유의한 차이를 나타냈다(p=0.002~0.026). 카파 통계량은 0.540에서 0.616으로 중등도의 일치도를 보였다(표 4).

비선형 모델 중 가장 우수한 모델(RBF-SVM)과 선형 모델 중에서 가장 우수한 모델(PLR)을 수신자 조작 특성 곡선(receiver operating characteristic curve)을 이용하여 서로 비

Table 2. Demographic characteristics of the included subjects

	Schizophrenia	Non-psychotic illness	Total	Statistics	p-value
Number	114	320	434		
Sex (n, %)				χ^2=1.58 (df=1)	0.208
Male	60 (52.6)	192 (60.0)	252 (58.1)		
Female	54 (47.4)	128 (40.0)	182 (41.9)		
Age (years)	31.1±11.6*	32.9±14.6	32.4±13.9	T=1.23 (df=432)	0.218
Education (years)	13.4±3.8	13.1±4.6	13.2±4.4	T=0.539 (df=432)	0.590

*: Mean±standard deviation

Table 3. The WAIS-IV subtests, index scores and their comparison between schizophrenia group and non-psychotic illness group

WAIS-IV	Schizophrenia	Non-psychotic illness	Cohen's d	T-value	df	p-value
Subtests						
Block design	6.91±3.72*	9.12±3.53	0.617	5.65	432	<0.001
Similarities	7.72±3.38	9.29±3.03	0.502	4.61	432	<0.001
Digit span	7.66±3.43	9.36±3.26	0.513	4.71	432	<0.001
Matrix reasoning	7.82±3.24	8.97±3.10	0.369	3.38	432	<0.001
Vocabulary	7.61±3.24	9.12±3.11	0.479	4.39	432	<0.001
Arithmetic	6.97±3.51	8.86±3.21	0.573	5.26	432	<0.001
Symbol search	4.38±2.84	7.72±3.67	0.966	8.84	432	<0.001
Visual puzzles	6.57±3.05	8.84±3.00	0.754	6.92	432	<0.001
Information	8.75±3.35	9.60±2.79	0.290	2.66	432	<0.001
Coding	4.45±2.75	7.39±3.49	0.889	8.15	432	<0.001
Index scores						
Verbal comprehension	89.3±16.5	96.8±14.9	0.486	4.46	432	<0.001
Perceptual reasoning	82.3±17.4	94.6±17.3	0.713	6.54	432	<0.001
Working memory	84.9±18.4	95.4±17.0	0.605	5.55	432	<0.001
Processing speed	70.3±16.6	88.1±18.1	1.003	9.19	432	<0.001
Total IQ	77.8±17.2	91.5±17.1	0.798	7.32	432	<0.001

*: Mean±standard deviation. WAIS-IV: Wechsler Adult Intelligence Scale fourth edition, IQ: Intelligence quotient

Table 4. The performance of the diagnostic models built by different machine learning algorithms

Algorithm type	Machine learning algorithms	Baseline model		Proper model	
		Accuracy (p-value)	Kappa	Accuracy (p-value)	Kappa
Linear models	Logistic regression	71.5% (0.760)	0.141	77.7% (0.185)	0.369
	Penalized logistic regression			78.5% (0.135)	0.397
	Linear SVM			76.2% (0.313)	0.325
	LDA			76.9% (0.245)	0.311
Nonlinear models	Radial basis function kernel SVM			84.6% (0.002)	0.616
	Random forest			83.1% (0.009)	0.545
	Gradient boosting			81.5% (0.026)	0.540

SVM: Support vector machine, LDA: Linear discriminant analysis

교하였다. RBF-SVM의 곡선 아래 면적(area under the curve, 이하 AUC)은 0.899, PLR의 AUC는 0.753으로 비선형 모델이 더 우수하였다(그림 1).

고 찰

본 연구에서는 비정신증 환자로부터 조현병 환자를 감별진단하는 과제에 머신 러닝 기법을 적용해보았다. 예측에 사용할 자료로는 WAIS-IV로부터 얻어지는 각종 지표를 이용하였다. 일반 통계 기법에 비하여, 비선형/비모수 알고리즘이 더 유용하리라는 가설을 세웠고, 이를 입증하기 위해 선형 알고리즘과 비선형 알고리즘을 사용하는 모델의 정확도를 비교하였다. 그 결과 가장 높은 예측 정확도는 비선형 알고리즘 중의 하나인 RBF-SVM에서 얻어졌으며, 정확도는 84.6%로서 기저 모델(71.5%) 및 가장 우수한 선형 모델인 PLR(78.5%)보다 우수하였다. 이를 통해 조현병의 감별진단이라는 패턴인식 문제에 있어서 비선형 알고리즘을 이용한 머신 러닝이 유용할 수 있음을 확인하였다.

임상에서 조현병 진단을 내릴 때는 WAIS 결과를 그다지 중요하게 고려하지 않는다.[31] 그 원인으로는 먼저 통계적으로는 차이가 나지만, 효과 크기가 작아 정상 대조군과 상당히 겹친다는 점을 들 수 있다.[32] 두 번째, 인지 기능은 매우 이질적인 기능들의 집합으로 단일 변수로 요약되기 어렵다는 점을 들 수 있다. 조현병의 인지 기능 저하는 다양한 변수에서 드러나며, 특이한 패턴을 찾기 어렵다.[18,19,33] 따라서 변수 하나하나에 대한 단일 검증으로는 조현병을 기타 질환으로부터 구분할 수 없다.[34] 예를 들어 조현병 환자와 정상 대조군을 비교한 대규모 메타분석에서는 기호 쓰기가 가장 큰 효과 크기(Hedges's g=1.59)를 보였는데, 이때의 효과 크기를 비중복 백분위로 환산하면 73%에 지나지 않았다.[35] 즉 기대

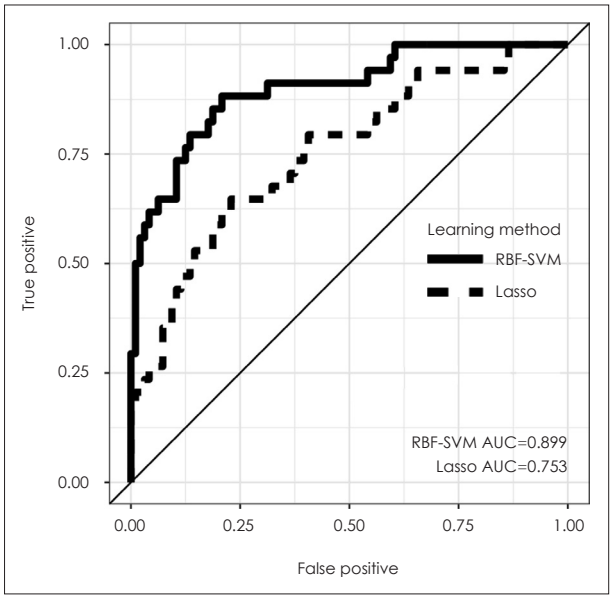

Fig. 1. Receiver operation curve depicting the model performances of the best nonlinear model (radial basis function kernel support vector machine) and the best linear model (penalized logistic regression). RBF-SVM : Radial basis function kernel support vector machine, Lasso : Penalized logistic regression, AUC : Area under the curve.

할 수 있는 최대 예측 정확도가 73%라는 의미이다. 본 연구의 최종 모델 정확도가 이보다 높다는 사실은 여러 변수를 동시에 고려하여 패턴을 찾는 머신 러닝의 장점을 드러내어 준다. 머신 러닝은 변수들 간의 다차원적 상호작용을 동시에 고려하여 패턴을 찾아내기 때문에, 다수의 변수를 하나로 요약할 필요도 없고 개개 변수를 이용한 단일 검증의 한계에 묶이지도 않는다.[36] 이 때문에 머신 러닝은 이미 뇌영상 연구에서 얻어지는 방대한 데이터를 처리하는데 유용한 방법론으로 자리잡았으며, 우울증 효과를 예측하는 등의 임상연구 데이터에도 적용되어 긍정적인 성과를 내고 있다.[37]

세 번째, 기분 장애에서도 인지 기능 저하가 일관되게 관찰된다는 점은 더욱 더 감별의 어려움을 가중시킨다.[15,20,38,39] 본 연구에서와 같이, 감별의 대상이 정상 대조군이 아니라 기타 정신질환군이라면, 기분 장애 환자가 감별대상이 되면서 정확도는 떨어질 수밖에 없다. 본 연구에 포함된 비정신증 환자 중 약 절반 가량은 우울증과 양극성 장애(각각 44.1%와 5.6%) 즉 기분 장애 환자였다. 기분 장애군과 기분 장애를 제외한 기타 비정신증군 사이에 WAIS 점수를 비교하면, 어휘, 동형찾기, 퍼즐, 지각추론, 처리속도, 전체 지능 지수가 유의한 차이를 보였다. 이렇듯 인지 기능이 저하된 기분 장애를 포함한 상태에서 감별하는 과제였음에도 불구하고, 본 연구의 예측 정확도는 84.6%에 달했다. 이는 머신 러닝이 정상 대조군과의 구분뿐만 아니라 기분 장애와 조현병을 구분하는 데도 어느 정도 도움을 줄 수 있음을 시사한다.

본 연구에서 비록 비선형/비모수 알고리즘이 더 우수한 예측 정확도를 나타낸다는 것을 확인할 수 있었지만, 비선형 알고리즘은 나름대로 단점이 있다. 이론적으로 비선형/비모수 모델은 입력과 출력 사이의 복잡한 함수적 관계를 임의의 정확도로 추정할 수 있다.[40] 그러나 비선형 모델이 지나치게 복잡해지면 과적합(overfitting)의 위험이 증가하여 새로운 데이터에 대한 일반화 능력이 떨어진다.[41] 특히 본 연구에서처럼 표본 크기가 작은 경우 과적합의 위험은 더욱 높아진다. 비록 본 연구에서 보고한 모델의 적합성은 학습에 사용되지 않은 테스트 데이터를 사용하여 얻어진 것이지만, 작은 표본에 복잡한 모델을 사용함으로써 정확도가 부풀려졌을 가능성을 감안해야 한다.

더군다나 본 연구는 단일 기관의 자료를 대상으로 하였기 때문에 그 적용범위에 한계가 있다. 모델의 일반화 가능성은 훈련 데이터의 포괄성에 의해 크게 좌우된다. 따라서 단일 기관의 자료를 통해 훈련된 모델은 다른 기관의 환자 집단에는 적용되지 않을 수 있다. 유용한 모델을 위해서는 다양한 상황에서 수집된 자료가 필요하다. 또 다른 한계점은 특이성이 낮다는 점이다. 본 연구의 최종 모델에서 특이도(73.5%)는 민감도(87.5%)보다 낮았고, 이는 비정신증으로 예측되는 환자 중 상당수가 실제로는 조현병이었다는 것을 의미한다. 다수의 조현병 환자가 표준 신경 심리 검사[42,43]에서 정상 소견을 보이며 이러한 조현병 환자의 비율은 16~45%로 추정된다.[44] 따라서 신경 인지 검사 결과만을 기반으로 하여 조현병을 진단하는 것은 이론적인 면에서도 극복할 수 없는 한계가 있다. 부인할 바 없이 최종 진단은 면밀한 면담과 포괄적인 정보 수집을 기반으로 하여 내려져야 한다. 그렇지만 이것이 감별진단에 참고할 수 있는 도구로써 WAIS-IV의 가능성을 부정하는 것은 아니다.

결 론

본 연구에서는 WAIS-IV 자료에 머신 러닝 알고리즘을 적용하여 조현병을 기타 비정신증 진단군으로부터 감별하는 진단 모델을 구축하였다. 얻어진 모델의 성능을 고려하였을 때, WAIS-IV가 조현병 감별진단을 도울 수 있다는 가능성을 확인하였으며, 동시에 정보를 효과적으로 이용하기 위해선 다수의 변수가 이루는 복잡한 패턴을 추출하는 비선형/비모수 머신 러닝 기법이 유리하다는 것을 알 수 있었다. 그러나 진단 모델의 성능과 적용범위를 높여 실제 임상에서 감별진단에 참고할만한 도구로 사용될 수 있으려면, 다기관으로부터 얻어진 대규모의 포괄적인 자료를 요할 것이다.

중심 단어 : 머신 러닝 · 조현병 · 웩슬러 지능검사 4판 · 신경인지기능 · 진단지원시스템.

Conflicts of Interest

The authors have no financial conflicts of interest.

REFERENCES

1) Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction, 2nd ed. New York: Springer Science & Business Media;2009.
2) Clarke B, Fokoue E, Zhang HH. Principles and theory for data mining and machine learning. New York: Springer Science & Business Media;2009.
3) Zhou ZH. Ensemble methods: foundations and algorithms. Boca Raton: CRC Press;2012.
4) Mikolas P, Melicher T, Skoch A, Matejka M, Slovakova A, Bakstein E, et al. Connectivity of the anterior insula differentiates participants with first-episode schizophrenia spectrum disorders from controls: a machine-learning study. Psychol Med 2016;46:2695-2704.
5) Deo RC. Machine learning in medicine. Circulation 2015;132:1920-1930.
6) Ravan M, Hasey G, Reilly JP, MacCrimmon D, Khodayari-Rostamabad A. A machine learning approach using auditory odd-ball responses to investigate the effect of Clozapine therapy. Clin Neurophysiol 2015;126:721-730.
7) Shim M, Hwang HJ, Kim DW, Lee SH, Im CH. Machine-learning-based diagnosis of schizophrenia using combined sensor-level and source-level EEG features. Schizophr Res 2016;176:314-319.
8) Johannesen JK, Bi J, Jiang R, Kenney JG, Chen CA. Machine learning identification of EEG features predicting working memory performance in schizophrenia and healthy adults. Neuropsychiatr Electrophysiol 2016;2:3.
9) Koutsouleris N, Kahn R, Chekroud AM, Leucht S, Falkai P, Wobrock T, et al. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. Lancet Psychiatry 2016;3:935-946.
10) Weickert CS, Weickert TW, Pillai A, Buckley PF. Biomarkers in schizophrenia: a brief conceptual consideration. Dis Markers 2013;35:3-9.
11) Tomasik J, Schwarz E, Guest PC, Bahn S. Blood test for schizophrenia. Eur Arch Psychiatry Clin Neurosci 2012;262 Suppl 2:S79-S83.
12) Michel NM, Goldberg JO, Heinrichs RW, Miles AA, Ammari N, McDermid Vaz S. WAIS-IV profile of cognition in schizophrenia. Assessment 2013;20:462-473.
13) Schaefer J, Giangrande E, Weinberger DR, Dickinson D. The global cognitive impairment in schizophrenia: consistent over decades and around the world. Schizophr Res 2013;150:42-50.
14) Kahn RS, Keefe RS. Schizophrenia is a cognitive illness: time for a change in focus. JAMA Psychiatry 2013;70:1107-1112.
15) Bora E, Harrison BJ, Yücel M, Pantelis C. Cognitive impairment in euthymic major depressive disorder: a meta-analysis. Psychol Med 2013;43:2017-2026.
16) Seaton BE, Goldstein G, Allen DN. Sources of heterogeneity in schizophrenia: the role of neuropsychological functioning. Neuropsychol Rev 2001;11:45-67.
17) Keefe RS, Fenton WS. How should DSM-V criteria for schizophrenia include cognitive impairment? Schizophr Bull 2007;33:912-920.
18) Dickinson D, Iannone VN, Wilk CM, Gold JM. General and specific cognitive deficits in schizophrenia. Biol Psychiatry 2004;55:826-833.
19) Dickinson D, Ragland JD, Gold JM, Gur RC. General and specific cognitive deficits in schizophrenia: Goliath defeats David? Biol Psychiatry 2008;64:823-827.
20) Trivedi MH, Greer TL. Cognitive dysfunction in unipolar depression: implications for treatment. J Affect Disord 2014;152-154:19-27.
21) Iwabuchi SJ, Liddle PF, Palaniyappan L. Clinical utility of machine-learning approaches in schizophrenia: improving diagnostic confidence for translational neuroimaging. Front Psychiatry 2013;4:95.
22) Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015;349:255-260.
23) Wechsler D. The psychometric tradition: developing the wechsler adult intelligence scale. Contemp Educ Psychol 1981;6:82-85.
24) Sumiyoshi C, Uetsuki M, Suga M, Kasai K, Sumiyoshi T. Development of brief versions of the Wechsler Intelligence Scale for schizophrenia: considerations of the structure and predictability of intelligence. Psychiatry Res 2013;210:773-779.
25) Fujino H, Sumiyoshi C, Sumiyoshi T, Yasuda Y, Yamamori H, Ohi K, et al. Performance on the Wechsler Adult Intelligence Scale-III in Japanese patients with schizophrenia. Psychiatry Clin Neurosci 2014;68:534-541.
26) Heinrichs RW, Zakzanis KK. Neurocognitive deficit in schizophrenia: a quantitative review of the evidence. Neuropsychology 1998;12:426-445.
27) Hwang ST, Kim JH, Park KB, Choi JY, Hong SH, editors. Standardization of the K-WAIS-IV. Proceedings of Korean Psychological Association Annual Conference; 2012 Aug 24; Chuncheon, Korea. Seoul: Korean Psychological Association;2012.
28) McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22:276-282.
29) R Development Core Team. R: a language and environment for statistical computing. 3.2.4 ed. Vienna: R Foundation for Statistical Computing;2016.
30) Kuhn M. Caret: classification and regression training. R package version 6.0-68 ed 2016.
31) Harrison-Read P. IQ tests as aids to diagnosis and management in early schizophrenia. Adv Psychiatr Treat 2008;14:235-240.
32) Bora E, Yucel M, Pantelis C. Cognitive functioning in schizophrenia, schizoaffective disorder and affective psychoses: meta-analytic study. Br J Psychiatry 2009;195:475-482.
33) Keefe RS, Perkins DO, Gu H, Zipursky RB, Christensen BK, Lieberman JA. A longitudinal study of neurocognitive function in individuals at-risk for psychosis. Schizophr Res 2006;88:26-35.
34) Kitchen H, Rofail D, Heron L, Sacco P. Cognitive impairment associated with schizophrenia: a review of the humanistic burden. Adv Ther 2012;29:148-162.
35) Dickinson D, Ramsey ME, Gold JM. Overlooking the obvious: a meta-analytic comparison of digit symbol coding tasks and other cognitive measures in schizophrenia. Arch Gen Psychiatry 2007;64:532-542.
36) Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry 2016;3:243-250.
37) Chekroud AM, Gueorguieva R, Krumholz HM, Trivedi MH, Krystal JH, McCarthy G. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. JAMA Psychiatry 2017;74:370-378.
38) Manove E, Levy B. Cognitive impairment in bipolar disorder: an overview. Postgrad Med 2010;122:7-16.
39) Vieta E, Popovic D, Rosa AR, Solé B, Grande I, Frey BN, et al. The clinical implications of cognitive impairment and allostatic load in bipolar disorder. Eur Psychiatry 2013;28:21-29.
40) Zielesny A. From curve fitting to machine learning: an illustrative guide to scientific data analysis and computational intelligence. 2nd ed. Switzerland: Springer International Publishing;2016.
41) Han H, Jiang X. Overcome support vector machine diagnosis overfitting. Cancer Inform 2014;13(Suppl 1):145-158.
42) Holthausen EA, Wiersma D, Sitskoorn MM, Hijman R, Dingemans

PM, Schene AH, et al. Schizophrenic patients without neuropsychological deficits: subgroup, disease severity or cognitive compensation? Psychiatry Res 2002;112:1-11.
43) Allen DN, Goldstein G, Warnick E. A consideration of neuropsychologically normal schizophrenia. J Int Neuropsychol Soc 2003;9:56-63.
44) Reichenberg A, Harvey PD, Bowie CR, Mojtabai R, Rabinowitz J, Heaton RK, et al. Neuropsychological function and dysfunction in schizophrenia and psychotic affective disorders. Schizophr Bull 2009;35:1022-1029.