당뇨병 및 내분비질환 분야 머신러닝 활용

홍남기, 박혜정, 이유미

연세대학교 의과대학 내과학교실 내분비내과

Machine Learning Application in Diabetes and Endocrine Disorders

Namki Hong, Heajeong Park, Yumie Rhee

Division of Endocrinology and Metabolism, Department of Internal Medicine, Yonsei University College of Medicine, Seoul, Korea

Abstract

Recently, machine learning (ML) applications have received attention in diabetes and metabolism research. This review briefly provides the basic concepts of ML and specific topics in diabetes research. Exemplary studies are reviewed to provide an overview of the methodology, main findings, limitations, and future research directions for ML-based studies. Well-defined, testable study hypotheses that stem from unmet clinical needs are always the first prerequisite for successful deployment of an ML-based approach to clinical scene. The management of data quality with enough quantity and active collaboration with ML engineers can enhance the ML development process. The interpretable high-performance ML models beyond the black-box nature of some ML principles can be one of the future goals expected by ML and artificial intelligence in the diabetes research and clinical practice settings that is beyond hype. Most importantly, endocrinologists should play a central role as domain experts who have clinical expertise and scientific rigor, for properly generating, refining, analyzing, and interpreting data by successfully integrating ML models into clinical research.

Keywords: Artificial intelligence; Deep learning; Diabetes mellitus; Machine learning; Metabolism

Corresponding author: Namki Hong

Department of Internal Medicine, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea, E-mail: nkhong84@yuhs.ac

Received: Jul. 14, 2020; Accepted: Jul. 24, 2020

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © 2020 Korean Diabetes Association

서론

머신러닝(machine learning, ML)은 최근 병원 자료의 전산화, 임상자료 누적, 대규모 자료를 처리할 수 있는 컴퓨터의 저장 및 계산능력의 증가와 함께 많은 관심을 받으며 여러 임상 분야 연구와 진료에서 활용되고 있다. 당뇨병과 내분비질환 분야에서, 머신러닝과 연관되어 게재된 논문의 수는 PubMed 검색기준 1986년부터 2020년 1월까지 약 2,000건에 달하며, 최근 10년간 기하급수적인 증가추세를 보이고 있다. 본 종설에서는 머신러닝과 연관된 여러 개념에 대한 간략한 정의, 당뇨병과 내분비질환에서 머신러닝을 활용한 예시 연구들을 살펴보고, 추후 당뇨병 및 내분비질환 임상연구와 진료에 있어 머신러닝의 활용방안 및 발전가능성에 대해 논의하고자 한다.

머신러닝: 개념 소개

1. 인공지능, 머신러닝, 그리고 딥러닝

인공지능(artificial intelligence, AI), 머신러닝(ML), 딥러닝(deep learning)은 의학 분야에서 많은 경우 동시에 사용되거나 같은 의미로 혼용되고 있다. 미국식품의약국(U.S. Food and Drug Administration)은 인공지능을 존 맥커시의 정의에 따라 ‘지능이 있는 기계를 만드는 과학과 공학(the science and engineering of making intelligent machines, especially intelligent computer programs)’으로 정의하였다[1,2]. 일반적으로 지능은 자기인식, 경험적 지식 등 포괄적인 영역을 지칭하는 개념이지만, 그 중 학습과 추론은 현 시점에서 인공지능을 정의하는 가장 중요한 요소로 인지되고 있다[1]. 인공지능이 가장 포괄적인 상위 개념이라면, 머신러닝은 ‘데이터로부터 스스로 학습하여 분류나 예측 등의 수행능력을 개선할 수 있는 소프트웨어나 알고리즘을 디자인’하기 위한 인공지능의 한 분야로 정의될 수 있다. 따라서 머신러닝은 인공지능에 포함되는 개념이나, 인공지능이 꼭 머신러닝을 지칭하는 것은 아니다. 딥

러닝 혹은 심층신경망(deep neural network)은 머신러닝의 한 종류로, 여러 겹의 신경망을 쌓아 사람의 두뇌와 유사하게 설계한 알고리즘을 지칭한다. 2018년 개최된 3차 ML for Health workshop에 따르면, 전통적인 공학분야 엔지니어 및 연구자들은 본인들의 연구분야를 지칭할 때 머신러닝이라는 단어를 더 선호하였으나, 임상 의사들은 더 넓은 의미의 인공지능이라는 단어를 흔히 활용하였다. 연구 커뮤니티 간 용어활용 차이는 소통 장애 및 불필요한 오류를 유발할 가능성이 있어, 머신러닝 관련 용어의 정확한 정의 및 사용에 대한 공통의 논의가 추후 필요할 것으로 보인다[3].

2. 머신러닝 알고리즘과 평가지표

머신러닝 알고리즘은 크게 지도학습(supervised learning), 비지도학습(unsupervised learning), 준지도학습(semi-supervised learning), 강화학습(reinforcement learning) 4가지 범주로 나누어 볼 수 있다(Table 1) [4-7]. 지도학습은 머신러닝 알고리즘을 학습시키기 위해 데이터와 정답지(레이블, label)를 함께 필요로 한다. 지도학습의 목표는 기존 데이터와 레이블과의 관계를 바탕으로, 새로운 데이터가 들어왔을 때 기존 자료에 비추어 가장 적합한 예측값을 추론하는 것이다. 비지도학습은 레이블이 없는 데이터를 탐색하며, 구조나 패턴을 찾거나 차원을 축소하는(정보량이 상대적으로 적은 변수들을 배제하고 전체 데이터를 잘 설명하는 일부 변수들만 남기거나 설명력이 높은 변수를 새로 생성하여 전체 변수 수를 줄이는) 작업을 수행한다. 준지도학습은 지도학습과 비지도학습방식을 함께 활용하여, 대규모 데이터에서 일부 데이터에만 레이블이 제시되어 있을 경우 유용할 수 있다. 강화학습은 주어진 복잡한 환경에 대해 특정 행동을 취했을 때 주어지는 보상 혹은 손실에 대한 학습을 통해, 누적보상을 최대화하는 최적의 행동을 탐색하며, 이러한 학습은 실제 작업을 진행하며 발생하는 행동-반응 데이터를 통해 실시간으로 이루어진다. 해결하고자 하는 문제에 대해 어떤 머신러닝 알고리즘을 선택할 것인가에 대해 일부 참고 가능한 지침들이 있긴 하지

Table 1. Machine learning algorithms

Types of learning	Supervised learning	Semi-supervised learning	Reinforcement learning	Unsupervised learning
Concept	Learning a function that best approximates new input to the desired output based on a given relationship between the input and labeled output from the labeled dataset	A mixed approach of supervised and unsupervised learning applicable to a small amount of labeled data and a large amount of unlabeled data	Learning by maximizing the reward function based on the responses yielded by various actions to achieve arbitrary goals in a given unstructured or unknown environment	Finding structures or patterns in an unlabeled dataset
Common tasks	Regression, classification	Regression, classification	Taking actions to maximize the reward	Clustering, dimensionality reduction
Estimators	Naïve Bayesian, decision tree, support vector machine (SVM), neural network, logistic/ridge/linear regression, elastic net, etc.	Generative model, semi-supervised SVM, etc.	Q-learning, policy gradient, actor-critic, etc.	K-means, density-based spatial clustering of applications with noise, auto-encoders, deep Boltzmann machine, principal component analysis, locally linear embedding, etc.
Examples	Prediction of gestational diabetes according to biochemical test results based on simple features extracted from an electronic health records database [4]	The DeepHeart algorithm, which provides cardiovascular risk scores based on heart rate monitoring from popular wearable devices (Fitbit, Apple Watch, etc.) [6]	Determining the optimal insulin dose in patients with type 1 diabetes based on activity, hemoglobin A1c level, alcohol consumption status, and the previous insulin dose [5]	Identifying novel clusters or biomarkers based on various features collected by an unbiased multimodal approach, which finds differences in risks for certain diseases compared to other groups [7]

만(예시: “cheat-sheet” for initial ML estimators, scikit-learn; https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html), 대부분의 경우 적절한 머신러닝 알고리즘의 선택은 데이터의 양, 구조, 문제 혹은 가설의 특성, 기존 지식, 연구자의 경험과 직관, 학습된 알고리즘의 성능 등을 고려하여 자료를 탐색하고 모델을 구축하는 반복적인 과정을 통해 이루어진다(Fig. 1). 이를 위해 해결하고자 하는 문제 및 데이터의 특성에 따라 머신러닝 모델에 대한 최적 평가지표가 신중히 결정되어야 한다(Table 2) [8-10].

당뇨병 및 내분비질환 머신러닝 적용 현황

최근 5년간 611개 영문 논문(인간 대상, 원저만 포함)의 제목을 분석하여 전체 2,155 단어 중 가장 빈번하게 사용된 30개 단어를 추출하였을 때, 질환 중에서는 diabetes나 diabetic이 가장 흔하게 등장하였으며(52%), retinopathy (14%), thyroid (14%)가 뒤를 이었다. 머신러닝 목적으로는 detection, classification, identification, diagnosis 등 진단 관련 단어가 40%로 가장 많았으며, risk prediction 혹은 prediction이 31%로 차순위, 그 외 segmentation (5%),

Fig. 1. A brief workflow of machine learning-based medical research.

Table 2. Commonly used machine learning performance metrics for binary or multiclass classifications [8-10]

Metrics	Concepts and equations
Accuracy	• $(TP + TN) / (TP + FP + FN + TN)$
Precision (= positive predictive value)	• $TP / (TP + FP)$
Recall (= sensitivity, true positive rate)	• $TP / (TP + FN)$
F1-score	• The harmonic mean of precision and recall • $2 \times (\text{recall} \times \text{precision}) / (\text{recall} + \text{precision})$
AUROC	• The area under the receiver operating characteristic curve (plotting TPR against FPR) • Higher AUROC, close to 1 = better classifier
AUPRC	• The area under the precision-recall curve (plotting precision against recall) • Higher AUPRC, close to 1 = better classifier • May have an advantage over AUROC when comparing the performance of models in an imbalanced dataset [10]

TP, true positive; TN, true negative; FP, false positive; FN, false negative; TPR, true positive rate; FPR, false positive rate; AUROC, area under the receiver operating characteristic curve.

혹은 bioinformatics (7%) 등이 있었다. 이 중 1) 지도, 비지도, 준지도, 혹은 강화학습 활용 사례를 제공할 수 있으며, 2) 최근 3년 이내 발표된 당뇨병 연관분야 연구 6편을

검토하였다. 해당 연구 세부 내용은 Table 3에 제시하였다 [4,5,11-14].

Table 3. Summary of recent studies related to machine learning applications in diabetes research

Task	Study (disease field)	Study subjects	Design and method	Key finding and limitation
Screening and diagnosis	Artzi et al., 2020 [4]	<ul style="list-style-type: none"> - Retrospective nationwide electronic health record data of 588,622 pregnancies from 368,351 women between 2010 to 2017 in Israel including data of demographics, anthropometrics, laboratory tests, diagnoses, and pharmaceuticals - Internal validation set (n = 137,220; with geo-temporal difference) 	<ul style="list-style-type: none"> - Aim: to establish an ML model to improve the prediction of gestational diabetes based on electronic health record vs. a conventional screening tool - Reference labels: gestational diabetes diagnosis by a two-step approach (glucose challenge test and oral glucose tolerance test at 24~28 weeks of gestation) - Comparator: National Institute of Health seven-item questionnaire - Methods: supervised learning; gradient boosting model 	<p>Key implications</p> <ul style="list-style-type: none"> - ML was useful in developing a simple nine-question model in self-reportable format from the large electronic health record dataset, which outperformed the current standard screening tool (AUROC 0.80 vs. 0.68). - May facilitate early-stage interventions for women at high risk for gestational diabetes - May aid construction of a selective, cost-effective screening approach according to predicted gestational diabetes risk instead of the current universal screening approach <p>Limitations</p> <ul style="list-style-type: none"> - Inherent bias from retrospective electronic health record data review - Performance might be different when based on actual self-reported surveys.
	Perakakis et al., 2019 [11]	<ul style="list-style-type: none"> - Serum samples of 49 healthy subjects and 31 patients with biopsy-proven NAFLD - Internal validation with three-fold cross-validation 	<ul style="list-style-type: none"> - Aim: to train models for the non-invasive diagnosis of NASH and liver fibrosis based on circulating lipids, glycans, fatty acids identified by liquid Chromatography with tandem mass spectrometry LC-MS/MS and biochemical parameters - Reference label: biopsy-proven NAFLD - Comparator: not applicable - Methods: supervised learning; one-vs-rest nonlinear support vector machine models with recursive feature elimination 	<p>Key implications</p> <ul style="list-style-type: none"> - The ML model including 20 features consisted of lipidomics, glycans, and adiponectin yielded high accuracy up to 90% in discriminating healthy individuals from patients with NAFLD and NASH. - May provide a low-risk cost-effective, non-invasive alternative method to liver biopsy. <p>Limitations</p> <ul style="list-style-type: none"> - Validation cohort was not available - Needs to be further validated in a different population.

Risk prediction Segar et al., 2019 [12]	<ul style="list-style-type: none"> - 8,756 Patients without heart failure at baseline from the ACCORD trial dataset (50% training set; 50% internal validation set; conducted between 1999 to 2009) - External validation set: 10,819 participants without prevalent heart failure from the ALLHAT trial 	<ul style="list-style-type: none"> - Aim: to develop an ML model to predict incident heart failure among patients with type 2 diabetes - Reference label: incident hospitalization or death due to heart failure (captured and adjudicated by two independent reviewer physicians during the trial) - Comparator: not applicable - Methods: supervised learning; random survival forest-based model 	<p>Key implications</p> <ul style="list-style-type: none"> - The ML-based models showed modest performance in prediction for incident heart failure among patients with type 2 diabetes in the external validation set (C-index 0.70 to 0.74). - Each 1-unit increment in the WATCH-DM score was associated with a 24% higher relative risk of heart failure within 5 years - Strength of analyzing a large number of participants from a well-phenotyped clinical trial population <p>Limitations</p> <ul style="list-style-type: none"> - Discrimination for heart failure with preserved ejection fraction was relatively low in the subgroup analysis. - Temporal changes of heart failure biomarkers and medications could not be reflected in the model. - Need to validate the model in lower-risk cohorts of individuals with type 2 diabetes
Basu et al., 2018 [13]	<ul style="list-style-type: none"> - 10,251 ACCORD trial participants aged 40 to 79 years with type 2 diabetes, HbA1c 7.5% or higher, or cardiovascular diseases or risk factors, those who randomized to target HbA1c < 6.0% (intensive) vs. 7.0% to -7.9% (standard group) 	<ul style="list-style-type: none"> - Aim: to identify subgroups with a heterogeneous treatment effect in response to intensive glycemic therapy - Reference label: treatment effect defined as the absolute difference in the all-cause mortality rate between the intensive and standard therapy groups - Comparator: not applicable - Methods: supervised learning; gradient forest analysis 	<p>Key implications</p> <ul style="list-style-type: none"> - Compared to 3.7% increased mortality by intensive vs. standard therapy in group 4, group 1 showed a 2.3% mortality reduction in the intensive therapy group (95% CI, -0.2% to 4.5%), which made the obvious contrast with the main result from the study. - Identified characteristics of patients who may have benefited from intensive glycemic therapy (younger individuals with relatively low hemoglycosylation index) - Offered an example to find clinically meaningful subgroups with heterogeneous treatment effects using data from randomized trials. <p>Limitations</p> <ul style="list-style-type: none"> - Post hoc analysis of a single trial that was conducted before the development of recent diabetes medications with cardiovascular benefits

	Oroojeni et al., 2019 [5]	<ul style="list-style-type: none"> - Medical records of 87 patients with type 1 diabetes from Mass General Hospital; data for each patient's visits over a 10-year period (training set) between 2003 to 2013; HbA1c, body mass index, activity level, alcohol usage status, insulin (Lantus) dose - External validation with 60 cases 	<ul style="list-style-type: none"> - Aim: to explore an effective reinforcement learning framework for determining the optimal long-acting insulin dose for patients with type 1 diabetes - Reference label: physician-prescribed insulin dose - Comparator: not applicable - Methods: reinforcement learning; Q-learning with reward function set from HbA1c status at the visit and change of HbA1c from the past visit 	Key implications
				<ul style="list-style-type: none"> - The physician-prescribed insulin dose was within the dosing interval recommended by the Q-learning algorithm in 88% of test cases. - A proof-of-concept study to provide clinical decision support for determining insulin dose in patients with type 1 diabetes, by applying reinforcement learning algorithm
Translational research	Liu et al., 2020 [14]	<ul style="list-style-type: none"> - 20 Drug-naive individuals with prediabetes (discovery cohort) - Determined exercise responders and non-responders after 12-week high-intensity exercise training - Collected pre- and post-exercise period feces to analyze gut microbiota profile - Internal validation with 10-fold cross-validation 	<ul style="list-style-type: none"> - Aim: to find an ML model for predicting exercise responsiveness determined from exercise-induced alterations in the gut microbiota - Reference label: responders defined as a decrease in the homeostatic model assessment of insulin resistance greater than two-fold technical error - Comparator: not applicable - Methods: supervised learning; random forest model 	Limitations
				<ul style="list-style-type: none"> - Limited by omitting lifestyle information regarding diet, stress, and medication adherence - A relatively small training set - Only one type of insulin (Lantus) was examined in the model
				Key implications
				<ul style="list-style-type: none"> - The ML model identified 14 microbiome species and 15 metabolites from human feces were able to predict exercise responsiveness (AUROC 0.75 in the validation set). - Provide an example of applying ML principles to human-to-mice translational study based on microbiome dataset
				Limitations
				<ul style="list-style-type: none"> - Relatively small sample size - Limited to Chinese males only - Need further validation in different population set

ML, machine learning; AUROC, area under the receiver operating characteristic curve; NAFLD, non-alcoholic fatty liver disease; NASH, nonalcoholic steatohepatitis; CI, confidence interval.

선별검사 및 질환 진단

1. 선별검사 성능 개선

많은 연구자들이 머신러닝 알고리즘의 활용이 당뇨병 및 내분비질환의 선별검사 성능개선에 도움을 줄 수 있을지 연구해왔다. Artzi 등[4]은 전자의무기록 EHR 데이터베이스를 토대로 임신성당뇨병을 선별검사하는 알고리즘을 개

발하였다. 2010년부터 2017년도까지 이스라엘 36만여 명의 임신부로부터 수집된 58만건의 EHR 자료를 활용하여 임신성당뇨병 선별검사 머신러닝 알고리즘을 학습시켰다. 2,355개의 후보 변수 중, 연구자들은 gradient boosting model 기반으로 9개의 자가응답이 가능한 변수만을 활용한 모델을 만들었고 이는 기존 선별검사 방법인 24~28주 사이 당부하검사에 비해 더 조기에 시행할 수 있으면서도 좋은 성능을 보여주었다(area under the receiver operating

characteristics curve, 0.80 vs. 0.68).

2. 진단법 개선

잘 구축되고 검증된, 정확한 머신러닝 모델은 기존의 침습적 검사를 대체할 수 있는 가능성이 있다. 일례로 비알콜성 지방간(non-alcoholic fatty liver disease, NAFLD)은 전 세계적으로 급속히 유병률이 증가하고 있는 질환이지만, 여전히 조직생검이 확진의 gold standard로 되어 있어 진단과정의 부담을 초래한다. 한 연구에서 혈청으로부터 분석한 lipidomic, glycomic, liver fatty acid 데이터를 활용, support vector machine 기반 NAFLD 진단 알고리즘을 제시하였다[11]. 간섬유화 존재 여부에 대하여, 해당 모델은 10개의 단순한 lipid species를 활용한 모델로도 98%에 달하는 정확도를 보여주어, lipidomics에 기반한 머신러닝 알고리즘이 간생검에 대한 대안일 수 있는 가능성을 제시하였다. 다만, 이러한 알고리즘은 개발에 활용된 데이터셋의 근본적인 조건(인종 등)에 국한되어 다른 인구집단에서의 추가 검증이 반드시 필요하며 해당 알고리즘의 경우 경증의 NAFLD에 대해서도 좋은 성능을 확보할 수 있을지에 대한 연구가 추가로 필요하다[15].

위험도 예측

1. 임상경과 예측

정확한 임상경과 예측은 개별화된 치료 및 관리를 가능하게 한다는 점에서 중요하다. WATCH-DM score는 제2형 당뇨병 환자에서 심부전 발생 위험도를 예측하는 머신러닝 모델로, 기존 잘 알려진 무작위대조군 연구인 ACCORD 연구자료에서 모델을 학습시키고 또 다른 대규모 무작위대조군 연구자료인 ALLHAT 연구자료에서 외부검증(external validation)을 시행하여, 신뢰할만한 모델의 성능을 보여주었다[12]. Random survival forest 기반 머신러닝 모델과 함께, 임상 의사에게 친숙한 점수 기반 계산표를 제시하여 여

러 환경에서 활용할 수 있는 위험도 예측 모델을 제시하여 머신러닝을 활용한 좋은 임상연구의 표본을 보여주고 있다.

2. 치료반응 예측

머신러닝 알고리즘을 통해 기존 연구방식으로 탐색하기 어려웠던 다양한 치료반응군을 개별화하여 예측할 수 있는 가능성이 있다. 연구자들은 ACCORD 연구자료를 검토하여, 기존 표준 혈당관리 대비 집중적인 혈당관리를 시행하였을 때 보이는 다양한 반응군을 재분석하였다. 비록 2008년도 발표된 ACCORD 연구의 결론은 집중혈당관리가 표준치료법에 비해 사망률을 증가시킨다는 것이었지만, 머신러닝 알고리즘을 통한 사후분석에서 연구자들은 집중혈당관리가 실제 사망률 개선에 유익했던 대상군을 찾을 수 있었으며, 상대적으로 사망률 증가에 영향을 준 고위험군은 전체 대상군 중에서 소수임을 제시하여 기존 연구결과를 재해석할 수 있는 가능성을 보여주었다[13]. 이 연구는 기존 대규모 무작위대조군 연구의 결과가 신중하게 해석될 필요가 있으며, 머신러닝을 활용하여 동일치료법에 대해 다양한 치료반응을 보일 수 있는 비균질한 집단을 찾아내고 이에 맞춘 개별화된 치료적 접근을 고려할 수 있음을 보여주었다. 한 연구는 강화학습을 통해 제1형 당뇨병 환자에서 운동량, 식이량 등 변화하는 환경조건을 반영하여 인슐린 필요량 결정을 도울 수 있는 가능성을 제시하였다[5]. 아직 성능이 충분하진 않지만, 이러한 연구결과들은 머신러닝이 치료반응 예측 및 이를 통한 개별화된 접근을 가능하게 하는 중요한 도구로 활용될 가능성을 시사한다.

중개연구

머신러닝 알고리즘은 현재 주목을 받고 있는 다중오믹스를 기반으로 한 임상-중개연구자료 분석에서, 수많은 변수들 중 주요 변수를 선택하고 모델을 효과적으로 구축하기 위한 주요한 도구로 활용될 수 있다. Liu 등[14]은 사람에서 운동 후 혈당 변화 반응성과 연관된 주요 장내미생물군

총 및 대사체를 탐색하기 위해 머신러닝 접근법을 활용하였다. 운동반응군과 비반응군은 장내세균총 및 대사체 패턴에 있어 유의한 차이를 보였으며, 분변이식술을 통해 쥐에게 운동반응군의 장내세균총 패턴을 만들어 주었을 때 운동 및 인슐린민감도 개선에 효과를 보였다. Random forest 알고리즘은 수천 개의 잠재적 후보 특성변수들 가운데 반응군과 비반응군을 가장 잘 감별하는 29개의 특성변수(장내미생물 14종, 15개 대사체) 조합을 찾았으며, 이러한 특성변수의 조합은 운동반응군과 비반응군을 예측하고 이에 따른 치료전략을 세울 수 있는 개별화된 접근을 위한 바이오마커로 기능할 가능성이 있다.

국내 머신러닝 활용 당뇨병 연구

국내 머신러닝 활용 의학연구는 딥러닝 기법의 발전과 함께 주로 영상의학 분야에서 활발하게 이루어졌고, 당뇨병 및 내분비질환 임상연구에서도 최근 적극적으로 머신러닝을 활용하려는 연구자들이 늘어나고 있다. 병원자료 기반으로 5년 내 당뇨발생률을 예측하는 모델을 제시한 연구에서는, 전자의무기록 기반 28개 변수를 추출하여 머신러닝 모델을 만들고 당뇨병 발생 여부를 예측하였다. 다만 예측력에 있어서 기존 예측모델을 상회하는 성능을 보여주지는 못하여, 기존 병원자료에서 얻을 수 있는 예측력의 한계를 넘기 위해서는 추후 환자의 생활패턴이나 식습관, 운동 등 병원 밖에서 수집될 수 있는 특성변수의 활용이 중요할 것으로 보인다[16]. 최근 활용이 증가하고 있는 연속혈당측정기(continuous glucose monitoring, CGM)에서 얻을 수 있는 혈당 변화 시계열 데이터는 머신러닝을 활용하기에 적절한 자료이다. 한 국내 연구에서, 연구진은 30분 이내 저혈당이 발생할 것을 예측하는 모델을 구축하였다. Random forest 모델은 area under the receiver-operating characteristics curve 0.966, 민감도 89.6%, 특이도 91.3%로 좋은 예측 능력을 보여 추후 CGM 및 인공췌장 개발, 고도화에 도움을 줄 가능성이 제시하였다[17].

결론

머신러닝의 활용은 데이터 활용, 해석에 있어 새로운 가능성을 제시하며 당뇨병 및 대사질환 연구에 도움을 줄 가능성이 있지만, 현재 의료데이터의 복잡성, 부정확성, 잠재적 오류를 고려할 때 대규모 자료의 확보와 머신러닝의 적용은 신중한 접근이 필요하다[18]. 정확한 문제 정의 혹은 가설 설정, 데이터 질(quality)에 대한 지속적인 관심, 임상 현장 및 미충족수요에 기반한 연구디자인 및 머신러닝 전문가 커뮤니티와의 적극적이고 투명한 협력이 머신러닝을 통한 당뇨병 및 대사질환 연구를 성공적으로 시행할 수 있는 중요한 요소가 될 수 있다[19]. 머신러닝을 당뇨병과 내분비질환 연구에 효과적으로 적용하여 임상적으로 유용한 결과를 이끌어내기 위해서는, 대규모, 고품질의 의료데이터의 구축 및 접근성 확보가 중요하다. 이를 위하여 병원자료, 지역사회코호트, 환자 레지스트리, 건강보험 청구자료, 개인 라이프로그 등 여러 자료들이 익명성을 유지한 채 연계된 다면데이터 구축, 정제 및 공통자료모델 등을 이용한 효율적인 다기관 연구 모델 수립이 필요하다. 점차 증가하는 머신러닝 기반 연구결과들을 정확하게 이해, 비판적 평가 후 임상에 도입하여 실제 환자들에게 도움을 줄 수 있도록 진료현장을 변화시키기 위해서는, 전문임상경험을 보유하고 있으며 임상현장의 미충족수요를 잘 인지하고 있는 당뇨병 및 대사질환 내분비 임상의학사의 역할이 필수적이다.

REFERENCES

1. McCarthy J. From here to human-level AI. *Artifi Intell* 2007;171:1174-82.
2. McCarthy J. What is artificial intelligence? Available from: <http://www-formal.stanford.edu/jmc/whatisai.html> (Accessed on 9th June, 2020).
3. Beaulieu-Jones B, Finlayson SG, Chivers C, Chen I, McDermott M, Kandola J, et al. Trends and focus of machine learning applications for health research. *JAMA*

- Netw Open 2019;2:e1914051.
4. Artzi NS, Shilo S, Hadar E, Rossman H, Barbash-Hazan S, Ben-Haroush A, et al. Prediction of gestational diabetes based on nationwide electronic health records. *Nat Med* 2020;26:71-6.
5. Oroojeni Mohammad Javad M, Agboola SO, Jethwani K, Zeid A, Kamarthi S. A reinforcement learning-based method for management of type 1 diabetes: exploratory study. *JMIR Diabetes* 2019;4:e12905.
6. Ballinger B, Hsieh J, Singh A, Sohoni N, Wang J, Tison GH, et al. DeepHeart: semi-supervised sequence learning for cardiovascular risk prediction. *CoRR* 2018;arXiv:1802.02511.
7. Shomorony I, Cirulli ET, Huang L, Napier LA, Heister RR, Hicks M, et al. An unsupervised learning approach to identify novel signatures of health and disease from multimodal data. *Genome Med* 2020;12:7.
8. Dinga R, Penninx BWJH, Veltman DJ, Schmaal L, Marquand AF. Beyond accuracy: measures for assessing machine learning models, pitfalls and guidelines. *bioRxiv* 2019;743138.
9. Handelman GS, Kok HK, Chandra RV, Razavi AH, Huang S, Brooks M, et al. Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *AJR Am J Roentgenol* 2019;212:38-43.
10. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;10:e0118432.
11. Perakakis N, Polyzos SA, Yazdani A, Sala-Vila A, Kountouras J, Anastasilakis AD, et al. Non-invasive diagnosis of non-alcoholic steatohepatitis and fibrosis with the use of omics and supervised learning: a proof of concept study. *Metabolism* 2019;101:154005.
12. Segar MW, Vaduganathan M, Patel KV, McGuire DK, Butler J, Fonarow GC, et al. Machine learning to predict the risk of incident heart failure hospitalization among patients with diabetes: the WATCH-DM risk score. *Diabetes Care* 2019;42:2298-306.
13. Basu S, Raghavan S, Wexler DJ, Berkowitz SA. Characteristics associated with decreased or increased mortality risk from glycemic therapy among patients with type 2 diabetes and high cardiovascular risk: machine learning analysis of the ACCORD trial. *Diabetes Care* 2018;41:604-12.
14. Liu Y, Wang Y, Ni Y, Cheung CKY, Lam KSL, Wang Y, et al. Gut microbiome fermentation determines the efficacy of exercise for diabetes prevention. *Cell Metab* 2020;31:77-91.e5.
15. Katsiki N, Gastaldelli A, Mikhailidis DP. Predictive models with the use of omics and supervised machine learning to diagnose non-alcoholic fatty liver disease: a "non-invasive alternative" to liver biopsy? *Metabolism* 2019;101:154010.
16. Choi BG, Rha SW, Kim SW, Kang JH, Park JY, Noh YK. Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks. *Yonsei Med J* 2019;60:191-9.
17. Seo W, Lee YB, Lee S, Jin SM, Park SM. A machine-learning approach to predict postprandial hypoglycemia. *BMC Med Inform Decis Mak* 2019;19:210.
18. Kim HS, Kim DJ, Yoon KH. Medical big data is not yet available: why we need realism rather than exaggeration. *Endocrinol Metab (Seoul)* 2019;34:349-54.
19. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44-56.