Journal of the Korea Academia-Industrial cooperation Society
Vol. 23, No. 7 pp. 184-191, 2022

https://doi.org/10.5762/KAIS.2022.23.7.184
ISSN 1975-4701 / eISSN 2288-4688

당뇨병성 콩팥병 예측의 기계학습 적용을 위한 분류기 알고리즘별 성능 비교에 대한 연구

박윤진, 강혜경*
중부대학교 간호학과

Performance Comparison of Various Classification Algorithms of Machine Learning Applications for Predicting Diabetic Nephropathy

Yoonjin Park, Hyekyung Kang*
Department of Nursing, Joongbu University

요 약 본 연구의 목적은 대용량 자료의 분석 및 예측에 용이한 머신러닝 기법을 다양하게 적용하여 당뇨병성 콩팥병 발생 영향요인 분석 및 예측을 위한 가장 적합한 기계분석 알고리즘을 찾고자 하는 것이다. 본 연구의 데이터는 질병관리본부에서 시행한 국민건강영양조사 2015년(제6기 3차)부터 2019년(제8기 1차)의 총 5개년 자료를 대상으로 분석하였으며 최종 분석 대상자는 2015년 548명, 2016년 626명, 2017년 598명, 2018년 575명, 2019년 607명이다. 총 eGFR 감소의 정량적 예측을 위해 kNN, Decision tree, LGBM, Voting, XGBoost의 5가지 분류기(Classification) 알고리즘을 검토하였다. 학습 및 예측 정확도를 수치적으로 평가하기 위한 지표로 평균제곱근오차(Root Mean Squre Erro, RMSE)와 결정 계수(R^2)를 활용하였다. 연구결과 XGBoost를 활용한 알고리즘의 RandomForest Regressor 기준으로 Hyper Parameter (gamma = 1.3, max_depth = 6)를 적용한 결과, 상관도(kendal)가 0.07 이상인 변수는 'sex', 'age', 흡연여부, 허리둘레, HDL-cholesterol, Hemoglobin, Hematocrit, Blood urea nitrogen, 백혈구, 요단백, 요중크레아티닌, 요나트륨으로 총 12개로 나타났으며, 기계학습의 결과 R2 Score는 0.752 최소 MAE는 0.231이었다. 본 연구를 바탕으로 당뇨합병증 위험도를 예측하는 새로운 예측 모델을 구성하여 머신러닝 모델을 웹서비스로 제공하여 실시간 건강관리에 활용함으로써 주요 질병 위험도 산출 및 합병증 예측을 통한 향후 효과적인 당뇨합병증 위험도 예측 서비스프로그램을 마련하는 기초 자료를 제공하고자 한다.

Abstract This study was undertaken to find the most suitable algorithm for analyzing and predicting the factors affecting the prevalence of diabetic nephropathy by using various machine learning techniques that are efficient for large datasets. We analyzed the data collected across five years through the National Health and Nutrition Examination Survey conducted by the Korea Centers for Disease Control and Prevention. The final analysis subjects included were 548 in 2015, 626 in 2016, 598 in 2017, 575 in 2018, and 607 in 2019. For quantitative prediction of eGFR reduction, five classification algorithms were reviewed: kNN, decision tree, LGBM, Voting, and XGBoost. To evaluate learning and prediction accuracy quantitatively, Root Mean Square Error and R^2 were used as indicators. Data were analyzed using XGBoost and the results were applied to a hyperparameter with the Random Forest Regressor as the standard of the algorithm, which showed the Kendall correlation to be greater than or equal to 0.07 from 12 factors. The result of machine learning showed an R^2 score of 0.752 and a minimum MAE of 0.231. Based on the findings of this study, we aim to construct a new prediction model for gauging the risk of developing complications of diabetes.

Keywords : Diabetes Mellitus, Renal Function, Kidney Disease, Machine Learning, XGBoost

본 논문은 한국연구재단 연구과제로 수행되었음(과제번호: 2021R1G1A1092286)
*Corresponding Author : Hyekyung Kang(Joongbu Univ.)
email: kanghk@joongbu.ac.kr
Received May 18, 2022 Revised June 20, 2022
Accepted July 7, 2022 Published July 31, 2022

1. 서론

1.1 연구의 필요성

만성신장질환(만성콩팥병)이란 사구체여과율에 관계없이 신장이 손상되거나 사구체여과율이 60mL/min 미만으로 3개월 이상 지속되는 상태를 말하며, 투석을 받는 말기신부전으로 진행하거나 심혈관계 질환 등 다양한 합병증으로 조기 사망에 이르는 심각한 질환이다. 만성신장질환은 2017년 기준 전세계적으로 9.1%의 유병률을 보였고, 1990년 이후 29.3%의 증가율을 나타냈으며[1], 국내에서도 2018년 남자 3.1%, 여자 10.7%가 증가된 가운데 특히 70세 이상 고령자는 15.1%라는 높은 유병률을 나타냈다[2]. 이러한 만성신장질환의 주요 원인은 당뇨병으로 당뇨병성 콩팥병은 신장의 비대, 사구체여과율 감소, 심한 단백뇨 및 말기신부전으로 진행된다고 알려져 있다[3]. 특히 제2형 당뇨병에서는 제1형 당뇨병과 달리 알부민뇨가 없을 수 있고, 고혈당, 인슐린 저항성, 단백뇨, 당화산물, 그리고 산화스트레스를 포함하는 여러가지 요인에 의해 사구체 여과율을 감소시킬 수 있어 조기 진단 및 예후를 향상시킬 수 있는 다양한 접근 방법이 필요하다[4].

당뇨병성 콩팥병을 예방하기 위하여 가장 중요한 것은 혈당 관리이다. 이는 “Legacy 효과” 라고 하며, 철저한 혈당조절이 고혈당으로 인한 불가역적 세포 내 손상(예: epigenetic damage)를 예방하는 것을 의미한다[5]. 하지만 당뇨병성 콩팥병을 예방하기 위해서는 혈당조절 이외의 당뇨로 인한 다른 합병증의 대한 조절도 필요하다. 예를 들어, 일부 선행연구에 의하면 2형 당뇨 환자에서 평균 수축기혈압이 10 mmHg 높아 질 때마다 미세 알부민뇨, 현성 알부민뇨, 사구체 여과율 감소(60 mL/min/1.73 m^2 이하), 또는 혈청 크레아티닌이 2배 이상 상승할 위험률이 15%씩 증가한다고 하였다[6]. 이 외에도 당뇨병의 대표적 합병증인 혈관질환으로 미세혈관 합병증과 대혈관합병증이 있다. 특히, 콩팥은 대부분 미세혈관으로 이루어진 장기인 만큼 미세혈관 합병증에 더욱 취약하며, 2020년 말기신부전의 원인 질환으로 당뇨병이 49.8%로 가장 높게 나타나 당뇨가 있는 투석환자가 당뇨가 없는 투석환자에 비해 사망률이 높다고 보고되었다[7].

따라서 당뇨병은 합병증에 관한 다각적인 분석이 필요하고 상당한 노력과 전문성이 요구되기에 복잡한 분석과정을 체계적으로 지원해줄 수 있는 다양한 분석 방법들을 적용할 필요가 있다. 하지만 당뇨병성 콩팥병의 위험예측요인은 대부분 전통적인 통계방법인 회귀분석과 구조모형분석 방법이 사용되거나 의사결정 나무(Decision Treee: 이하 DT)의 단일한 방법만을 적용하여 제한적으로 분석하는 경우가 많았다[8]. 하지만 최근에는 기계 분석을 통한 다양한 접근을 시도하는 연구가 국내외적으로 많이 진행되고 있다. 특히 국내에서는 머신러닝(Machine Learning: 이하 ML) 기법을 활용한 연구들은 인구 통계학적 정보와 의료기록 그리고 생활 습관을 통한 합병증 예측, 유전체, 영상기록, 의료기록을 통한 알즈하이머 치매 예측, 치매의 위험요소인 biomarker를 통한 예측 연구 등에 다양하게 적용될 수 있다[8]. 국외 연구에서는 최근에 ML 및 데이터마이닝(data mining)은 의학/생물학분야에서 질병 예측 및 식별을 위한 연구에 널리 사용되고 있다[9-11]. 예를 들어, 여러 ML을 기반으로 인체계측정보를 이용한 serum high-density(HDL) lipoprotein 콜레스테롤과 low -density lipoprotein (LDL) 콜레스테롤 예측 연구가 수행되어졌으며[10], 고중성지방혈증 예측 모델에 관한 연구도 보고되었다[12,13]. 이러한 연구들은 최근 국내외적으로 Deep learning, 인공지능 (Artificial Intelligence)을 기반으로 한 질병 예측 및 식별, 나아가 진단 연구로까지 진행되고 있다[14]. 하지만 당뇨성 질환은 망막 변성에 관련한 검사 방법에 국한되었고, 당뇨병성 콩팥병증의 합병증에 관하여 ML을 활용한 예측 연구는 시도되지 않았다.

따라서, 본 연구에서는 대용량 자료의 분석 및 예측에 용이한 ML 기법을 다양하게 적용하여 합병증 발생 영향요인 분석 및 예측을 위한 ML 기법의 적용 가능성과 그에 따른 문제점 및 해결방안을 함께 살펴보고, 당뇨병성 콩팥병을 예방하기 위한 가장 적합한 기계분석 알고리즘을 찾고자 한다. 뿐만아니라 이를 통하여 당뇨병성 콩팥병의 예측 모델을 개발하여 공중보건과 건강 예측 분야에서 활용할 수 있는 프로그램의 기초자료를 제공하고자 한다.

1.2 데이터 셋

본 연구는 질병관리본부에서 시행한 국민건강영양조사 2015년(제6기 3차)부터 2019년(제8기 1차)의 총 5개년 자료를 대상으로 분석하였다. 국민 건강 영양조사의 내용은 건강 설문과 이동검진으로 진행되며, 매년 192개로 표본 조사구 내에서 군대, 교도소 등의 시설 및 외국인 가구 등을 제외하였다. 조사대상자는 2015년 7,380명, 2016년 8,150명, 2017년 8,127명, 2018년 7,992명, 2019년 8,110명이었고, 이중 당뇨병 진단을 받은 만 18세 이상 성인 중 혈중 크레아티닌 검사를 시행한 대상

자를 분석대상으로 하였으며, 최종 분석 대상자는 2015년 548명, 2016년 626명, 2017년 598명, 2018년 575명, 2019년 607명이다.

1.3 데이터 선별

본 연구에서는 당뇨병성 콩팥병의 원인을 파악하기 위한 변수로 혈중 크레아티닌의 상승 요인을 분석하였다. 콩팥의 기능은 미국 National Kidney Foundation의 Kidney Disease Improving Global Outcomes (KDIGO) 지침에 의하면 신기능 정상군(eGFR ≥90mL/min), 신기능 경도 감소군(60mL/min≦ eGFR 〉90mL/min), 신기능 경중등도군(45mL/min≦ eGFR 〉60mL/ min), 신기능 중등중증군(30mL/min≦ eGFR 〉45 mL/min), 신기능 중증 감소군(eGFR 〈30mL/min) 등 5그룹으로 분류하며[15,16], 대한신장학회에서는 1단계 정상군(eGFR ≥90mL/min), 2단계(60mL/min≦ eGFR 〉90mL/min), 3단계(30mL/ min≦ eGFR 〉59mL/min), 4단계(15mL/min≦ eGFR 〉30mL/min), 5단계(eGFR 〈1mL/min)로 구분하고 3단계부터 신부전 단계로 정의하였다[17]. 신장의 기능은 Estimated Glomerular Filtration Rate(eGFR)을 활용하며, 본 연구에서는 혈중크레아티닌 결과치와 환자의 나이, 성별 및 인종 정보를 바탕으로 $eGFR_{MDRD}$ 계산식을 이용하여 계산하였다 ($eGFR_{MDRD}$(mL/min)= 186 × (Serum creatinine)-1.154×(age) - 0.203× 0.742 [if female]) [17].

당뇨병성 콩팥병 예측을 위해 본 연구에서는 의사로부터 당뇨병 진단을 받은 당뇨환자를 대상자로 기준 하였고, 조사 당시 혈중 크레아티닌을 검사한 대상자만 포함을 시켰으며, 변수 선별 단계에서 문자열로 표기된 변수와 결측치가 50% 이상인 변수는 제외하였다. 기본 설문 정보는 지역, 성별, 나이, 소득, 교육정도, 직업 등의 일반적인 기본 설문정보와 혈당, 당화혈색소(HBA1C), 혈액요소질소(BUN), 간수치(SGOT, SGPT), 콜레스테롤, 인슐린 등 혈액검사를 통해 분석한 수치와 소변산도, 요비중(Urine-SG) 등 소변 검사를 분석한 수치, 그리고 혈압, BMI, 허리둘레 등 신체 계측치 등의 건강상태 진단 정보 및 스트레스 관련 정보를 포함하여 아래 표와 같이 총 41개의 변수를 선별하였다. 데이터 선별 및 정제 기준은 당뇨병성 콩팥병과 인과관계가 있는 데이터를 선별 및 정제하였으며, 기계학습 모델에 적용하기 위해 정제된 변수를 수치화 하였다. 또한 누락된 데이터의 정제를 위해 선별된 변수 중 문자열, 중복사항, 미응답 항목 등을 수치화하여 적용하였다. 선별된 데이터 변수는 아래 표와 같다(Table 1).

Table 1. Data Attributes and Format at the First analysis

Attribute	Data Format
Age	Nominal (Female, male)
House income	Nominal (1.2,3,4)
Education	Nominal (1.2,3,4)
Occupation	Nominal (1,2,3,4,5,6,7)
Hypertension	Nominal (N, Y)
Dyslipidemia	Nominal (N, Y)
Myocardial infarction	Nominal (N, Y)
Angina	Nominal (N, Y)
Thyroid disease	Nominal (N, Y)
Depression	Nominal (N, Y)
Blood pressure	Numeric
Waist circumference	Numeric
Body mass index	Numeric
Blood sugar test	Numeric
Blood test	
Insulin	Numeric
HbA1C	Numeric
Total Cholesterol	Numeric
HDL cholesterol	Numeric
LDL Cholesterol	Numeric
SGOT/SGPT	Numeric
SGPT	Numeric
Hemoglobin	Numeric
Hematocrit	Numeric
Blood urea nitrogen	Numeric
Creatinine	Numeric
White Blood Cell	Numeric
Red Blood Cell	Numeric
C-reactive protein	Numeric
Thyroid stimulation hormone	Numeric
Thyroid hormones T3	Numeric
Thyroid hormones T4	Numeric
Urine Test	
Uric acid	Numeric
Nitrate of urea	Numeric
Specific gravity of urine	Numeric
Protein	Numeric
Glucose	Numeric
Ketone	Numeric
Bilirubin	Numeric
Cotinine	Numeric
Creatinine	Numeric
Sodium, iodine	Numeric
Iodine	Numeric

1.4 기계학습 알고리즘

1.4.1 분류알고리즘

본 연구에서 eGFR 감소의 정량적 예측을 위해 kNN, Decision tree, LGBM, Voting, XGBoost의 5가지 분류기(Classification) 알고리즘을 검토하였다. 이는 변수간의 관련성과 목표변수의 영향인자 파악. 목표변수의 예측, 주요 변수 예측을 목적으로 하는 분석방법으로 비교적 큰 테이블을 다루기 쉽고, 예측과 분류에 다양하게 활용할 수 있는 장점을 가진 분석 방법으로 알려져 있다[18].

k-NN(k-nearest neighbors)은 사례기반학습(Instance-based Learning) 알고리즘으로 임의의 객체(Instance)가 비슷한 특성을 가진 다른 객체들과 얼마나 근접하게 위치하여 있는지에 따라 특성이 결정된다. 예를 들면, 아래 그림에서 총 8개의 "+, -"데이터가 있는 경우, k를 2, 5, 8로 결정함에 따라 예측 결과는 달라질 수 있다. 분류의 정확성을 결정하기 위한 k-NN의 Hyper parameter는 탐색할 이웃의 수(k)와 거리 측정 방법이 있다. k가 작을 경우 모델이 과대적합(Overfitting)되고, 반대로 k가 클 경우 모델이 과소적합(Underfitting)되는 경향이 있다[19].

Decision tree은 통계학, 데이터 마이닝, 기계 학습에서 사용하는 예측 모델링 방법 중 하나로서 어떤 항목에 대한 관측값과 목표값을 연결시켜주는 예측 모델로서 결정 트리를 사용하며, 회귀 트리 분석은 예측된 결과로 특정 의미를 지니는 실수값을 출력할 수 있다. 즉 의사결정트리는 선택과 결과를 가지의 형태로 나타내고, node는 분류할 그룹의 속성이며, Branches는 node가 사용할 수 없는 값을 나타낸다[19].

Vorting은 Bagging과 투표방식이라는 점에서 유사하지만, Bagging은 같은 알고리즘 내에서 다른 sample 조합을 사용하는 반면, Voting은 다른 알고리즘 model을 조합해서 사용한다[20].

또한, LGBM(Light Gradient Boost Machine)은 최대 손실 값을 갖는 부분에서 리프 중심 트리 분할(Leaf Wise)을 함으로써 정확도가 높고, 기계학습을 수행하는 시간이 적다는 장점이 있으며[21], XGBoost는 의사결정나무 모델에 단순한 분류가 가능한 예측 모델들을 결합하여 더욱 강한 예측 모델을 만드는 부스팅 기법을 적용하여 정형 데이터를 예측할 때 훌륭한 성능을 보이는 알고리즘으로서, 회귀 모형 또는 분류 모형에서 활용 가능하다[22](Fig. 1).

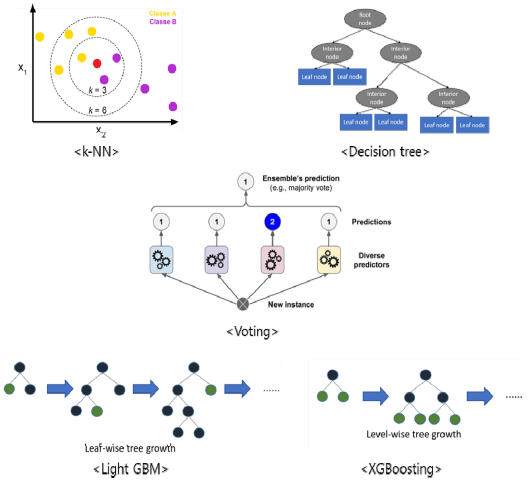

Fig. 1. Regression Algorithms for prediction of Diabetic Nephropathy

1.4.2 데이터 셋 적용

기계학습을 수행하기 위해서는 분석 데이터를 학습 데이터(Train Data)와 시험 데이터(Test Data)로 분류하여야 한다. 따라서 2절에서 전처리한 2010년부터 2019년까지의 데이터 중 일부를 학습데이터로, 나머지는 시험데이터로 나누어 기계학습을 수행하였다. 학습 정확도의 검증을 위해 학습 데이터의 20%를 랜덤으로 추출하여 검증 데이터(Validation Data)로 활용하였고, 교차검증(Cross Validation)을 통하여 모델을 학습하였다. 또한 시험 데이터는 학습된 모델의 예측 정확도를 확인하기 위해 사용되었다. 학습 및 예측 정확도를 수치적으로 평가하기 위한 지표로 평균제곱근오차(Root Mean Squre Erro, RMSE)와 결정 계수(R^2)를 활용하였다. R^2 값이 1에 가까울수록 모델의 정확도가 높음을 보여주고, RMSE는 실제 값과 예측 값의 차이를 나타내므로 작을수록 정확도가 높음을 의미한다[12].

2. 본론

2.1 당뇨병성 콩팥병 예측 성능

2.1.1 알고리즘 별 예측 성능

당뇨병성 콩팥병을 예측하기 위해서 선별된 데이터 셋을 대상으로 분류 알고리즘을 적용하여 각 알고리즘의 성능을 평가하였다. 알고리즘 별 예측모델의 성능평가 결과는 다음과 같다.

먼저, k-NN 알고리즘 적용 결과 혈중 크리아티닌의 최대 예측 정확도를 산정하기 위해 알고리즘의 Hyper Parameter(k=1)를 적용하였으며 최대 예측 정확도를 산정하기 위해서 변수간 상관도를 0.01% 간격으로 상승하여 변수를 선정하였다. 상관도(kendal)가 0.11 이상인 변수는 sex', 'age', ' Waist circumference', 'HDL cholesterol', ' Blood urea nitrogen', '요단백', '요나트륨'으로 총 7개로 나타났다. 기계학습의 결과 R^2 Score는 0.521 최소 MAE는 0.317이었다.

Decision Tree 혈중 크리아티닌의 최대 예측 정확도를 산정하기 위해 알고리즘의 Hyper Parameter (max_depth = 2)를 적용하였다. 상관도(kendal)가 0.13 이상인 변수는 'sex', 'HDL cholesterol', 'Blood urea nitrogen', '요나트륨'으로 총 4개로 나타났고, 기계학습의 결과 R^2 Score는 0.453 최소 MAE는 0.339이었다.

Voting 알고리즘 혈중 크리아티닌의 최대 예측 정확도를 산정하기 위해 알고리즘의 Linear Regression + Random Forest Regressor 결합하고 Hyper Parameter (n_estimators = 10, max_depth = 11)를 적용하였다. 상관도(kendal)가 0.00 이상인 변수는 sex, age, 가계수입, 교육정도, 직업, 음주여부, 흡연여부, 수축기혈압, 이완기혈압, 허리둘레, 체질량지수, 혈당, 당화혈색소, 총콜레스테롤, HDL- cholesterol, 중성지방, LDL-chlesterol, SGOT, SGPT, Hemoglobin, Hematocrit, Blood urea nitrogen, 백혈구, 적혈구, 혈소판, 고감도 C 반응단백검사, 요산도, 요아질산염, 요비중, 요단백, 요당, 요케톤, 요빌리루빈, 요잠혈, 유로빌리노겐, 요중크레아티닌, 요나트륨으로 총 38개의 변수로 나타났다. 기계학습의 결과 R^2 Score는 0.551 최소 MAE는 0.307이었다.

LGBM 알고리즘은 혈중 크리아티닌의 최대 예측 정확도를 산정하기 위해 알고리즘의 Random Forest Regressor 기준으로 Hyper Parameter (n_estimators = 10, max_depth = 11)를 적용하였다. 상관도(kendal)가 0.15 이상인 변수는 'sex', Blood urea nitrogen, 요중 크레아티닌, 요나트륨으로 총 4개로 나타났고, 기계학습의 결과 R^2 Score는 0.501 최소 MAE는 0.324이었다.

XGBoost 알고리즘은 혈중 크리아티닌의 최대 예측 정확도를 산정하기 위해 알고리즘의 Ran dom Forest Regressor 기준으로 Hyper Parameter (gamma = 1.3, max_depth = 6)를 적용하였다. 상관도(kendal)가 0.07 이상인 변수는 'sex', 'age', 흡연여부, 허리둘레, HDL-cholesterol, Hemoglobin, Hematocrit, Blood urea nitrogen, 백혈구, 요단백, 요중크레아티닌, 요나트륨으로 총 12개로 나타났다. 기계학습의 결과 R^2 Score는 0.752 최소 MAE는 0.231이었다. 또한, XGBoost를 활용한 분석 결과 변수 중요도는 BUN이 29.0으로 가장 높았으며, HDL-cholesterol 10.0, 나이 9.0, 허리둘레 8.0, 성별 8.0 순으로 높았다(Table 2)(Fig. 2).

Table 2. Prediction results by algorithm

Algorithm	Min. Value of MAE	R^2 Score
k-NN	0.317	0.521
Decision tree	0.339	0.453
Voting	0.307	0.551
LGBM	0.324	0.501
XGBoost	0.231	0.752

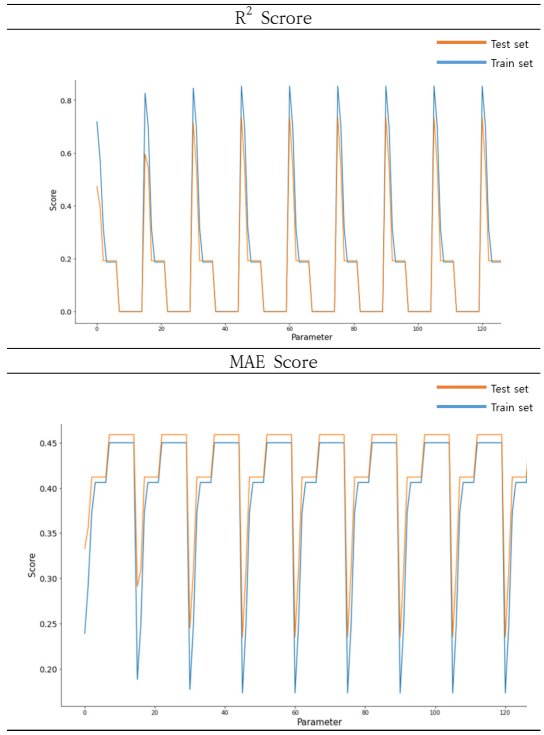

Fig. 2. Result of XGBoost

2.1.2 중요도

초기 모델을 구축하는데 41개의 변수가 이용되었으며 차원축소과정을 통해 최종 모델 개발에는 12개의 변수들이 상관관계가 있었으며 BUN, HDL-Cholesterol, 나이, 허리둘에, 성별, 백혈구, 유나트륨, 요단백, 요크레이티닌, 헤모글로빈 순으로 높은 중요도를 나타냈다(Fig. 3).

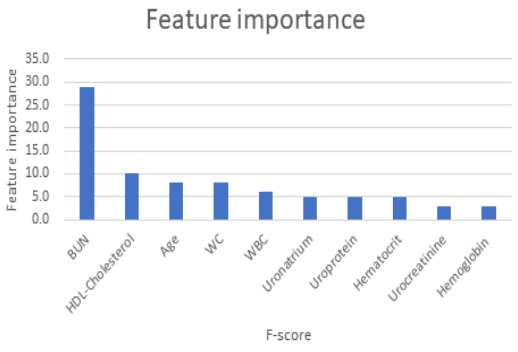

Fig. 3. Importance of variables used in XGBoost

3. 논의

본 연구는 당뇨병성 콩팥병 예측의 기계학습 적용을 위한 분류기 알고리즘 별 성능 비교에 대한 연구이다. 본 연구에 의하면 eGFR을 계산하기 위한 혈중 크리아티닌 수치 예측결과 'XGBoost Regression 모델'의 정확도가 가장 높았다. 선행연구에서도 XGBoost를 활용한 당뇨병 예측 알고리즘 연구에서도 평균 정확도 86%의 성능을 설명하였고[23], 폐경 여성에서 골다공증을 예측하는 모형에서도 모델의 성능이 가장 좋게 평가된 XGBoost 모델에서 10개 독립변수를 하나씩 축소하여 평가한 다른 분석방법보다 AUC가 가장 높게 나타났다[24]. 그리고 일부 연구에서는 Biomedical data 분석에 XGBoost를 사용하는 것이 다른 기계 분석을 이용하는 것 보다 우수하다고 평가하여[25] 본 연구와 유사한 결과를 나타내었다.

XGBoost는 앙상블 모델의 부스팅(Boosting)기법을 이용한 모델이다. 앙상블 모델은 여러 모델을 이용하여 학습을 하고, 모델의 예측결과를 평균하여 예측을 하는 방법으로 다수의 모델을 사용함에 따라 혼합 모델이라고도 한다. 그리고 배깅 (Bagging) 기법과 부스팅 기법으로 분류되는데 배깅 기법은 다수의 모델을 병렬 형식으로 학습시키는 기법이며, 부스팅 기법은 약한 분류가 가능한 다수의 모델을 결합하여 강한 모델을 만드는 것으로 주어진 데이터를 약한 분류기를 통해서 학습한 후 학습된 결과에서 나타나는 오차를 또 다른 약한 분류기에서 학습시켜 오차를 줄여 나가는 것이다. 첫 번째 학습을 통해 생성된 모델의 오류를 줄이기 위해 데이터를 여러 모델을 순차적으로 줄여 나가며 최종 결과는 각 모델 결과를 합하여 구한다. 대표적인 부스팅으로는 AdaBoost, XGBoost가 있고, 본 연구에서는 부스팅 기법을 이용했으며, 그 중에 XGBoost를 사용하였다[26]. 이는 biomedical data가 다양하며 환자의 상황에 따라 오차가 있어 이를 줄여 나가며 기계학습을 하는 것이 중요하기 때문이다. 이러한 방법은 당뇨 환자를 대상으로 SVM, RF 등을 분류기로 사용한 연구에서의 정확도 71~74(%)보다 높게 나타났다[27].

본 연구에서 XGBoost를 사용한 Feature Importance를 통해 BUN, HDL, 연령, 허리둘레, 그리고 성별이 혈중 크리아티닌 수치 여부를 가르는 주요 변수로 제시됨에 따라 이 변수들과 당뇨병성 콩팥병의 관계를 살펴볼 필요가 있다.

혈액요소질소(BUN)는 혈중 크리아티닌과 함께 신장 기능을 확인하는 생리적 지표로, 본 연구에서 Feature Importance가 제일 높았던 만큼 당뇨병성 콩팥병 예측에 XGBoost 알고리즘 모형이 적합한 것으로 판단할 수 있다. 또한, 제2형 당뇨병 환자에서 HDL-콜레스테롤이 높을수록 당뇨병성 콩팥병의 발병 위험이 낮고, HDL-콜레스테롤이 감소되면 혈중 크레아티닌이 증가한다는 선행연구를 고려할 때 HDL-콜레스테롤과 당뇨병성 콩팥병의 관련이 있음을 XGBoost 알고리즘 모형을 적용한 본 결과에서도 확인할 수 있다[28]. 뿐만 아니라 본 연구의 Feature Importance에서 세 번째로 높게 나타난 연령을 살펴보면 콩팥은 일반적으로 나이가 들면서 크기가 작아져 콩팥으로 가는 혈류가 줄어 노폐물을 걸러내는 사구체 여과율이 계속 감소하게 되어 고령일수록 당뇨병과 같은 만성질환 시 콩팥 기능 감소가 더 빨리 진행되는 만큼 연령이 당뇨병성 콩팥병과 관계있다는 선행연구를[5,29] 지지한다. 이 외에도 복부비만의 지표가 되는 허리둘레 역시 본 연구에서 혈중 크리아티닌 수치와 관련 있는 주요 변수로 제시되었는데, 이는 복부비만이 당뇨병성 콩팥병 발병에 영향을 미친다는 메타분석 결과와도 일치한다[30]. 뿐만 아니라, 성별은 이미 선행연구를[5,29] 통해 당뇨병성 콩팥병의 감수성 인자로 알려진 만큼 본 연구에서의 당뇨병성 콩팥병 예측의 기계학습 적용을 위한 XGBoost 알고리즘 모형의 정확도는 신뢰할 수준으로 판단된다. 특히, XGBoost의 경우 IQR (Interquartile Range) 이상치를 처리하지 않고 분석하는 경우 신체의 다양성을 바탕으로 데이터셋의 정확도를 높일 수 있어[31] 당뇨병성 콩팥병에 영향을 주는 여러 가지 원인을 분석을 위해 더욱 유용하다.

본 연구에서는 국민건강영양조사 원시자료를 이용하여 당뇨병성 콩팥병 발생 예측 및 주요 위험요인을 선별하기 위해 XGBoost 모형을 제안하였다. 훈련자료를 이

용하여 모형을 구축하고 검증자료에서 모형들의 예측성능을 R^2를 이용하여 평가한 결과, XGBoost의 예측성능이 k-NN, Decision tree, LGBM, Voting 비하여 우수한 것으로 나타났다. 그러나 당뇨병성 콩팥병은 다양한 원인에 의하여 발생할 수 있는 만큼 유의미한 변수들을 찾는데 한계가 있다. 또한 질병 발생에 관한 연구는 비교적 긴 추적관찰 기간이 필요하나 본 연구에서 사용된 국민건강영양조사는 횡단면 자료(cross-sectional data)로서 시간 경과에 따른 변화를 살펴볼 수가 없어 질병 발생의 추이 및 건강위험 요인들 간의 인과관계에 대한 파악이 어렵다는 단점이 있다. 하지만 본 연구를 바탕으로 당뇨합병증 위험도를 예측하는 새로운 예측 모델을 구성하여 머신러닝 모델을 웹서비스로 제공하여 실시간 건강관리에 활용함으로써 주요 질병 위험도 산출 및 합병증 예측을 통하여 향후 효과적인 당뇨합병증 위험도 예측 서비스프로그램을 마련하는 기초자료를 제공하고자 한다.

References

[1] H. J. Seo, "Effects for Comorbidities of Chronic Kidney Disease on the Progression to End-stage Renal Disease", *Journal of Health Informatics and Statistics,* Vol.45, No.4, pp.356-364, 2021.
DOI: https://doi.org/10.21032/jhis.2020.45.4.356

[2] K. W. Oh, Y. J. Kim, S. H. Kweon, S. Y. Kim, S. H. Yun, et al., "Korea National Health and Nutrition Examination Survey, 20th anniversary: accomplishments and future directions", *Epidemiology and health*, Vol.43, e2021025, 2021
DOI: https://doi.org/10.4178/epih.e2021025

[3] Association AD, "11. Microvascular complications and foot care: standards of medical care in diabetes-2019", *Diabetes Care,* Vol.42(Suppl 1), S124-S138, 2019.
DOI: https://doi.org/10.2337/dc19-S011

[4] T. Yamazaki, I. Mimura, T. Tanaka, M. Nangaku, "Treatment of diabetic kidney disease: current and future", *Diabetes & Metabolism Journal,* Vol.45, No.1, pp.11-26, 2021.
DOI: https://doi.org/10.4093/dmj.2020.0217

[5] D. H. Yang, S. Y. Lee, "Diabetic kidney disease: seven questions", *Journal of the Korean Medical Association,* Vol.63, No.1, pp.6-13, 2020.
DOI: http://doi.org/10.5124/jkma.2020.63.1.6

[6] R. Retnakaran, C. A. Cull, K. I. Thorne, A. I. Adler, R. R. Holman, UKPDS Group, "Risk factors for renal dysfunction in type 2 diabetes: UK Prospective Diabetes Study 74", *diabete,* Vol.55, No.6, pp.1832-1839, 2006.
DOI: https://doi.org/10.2337/db05-1620

[7] Y. A. Hong, T. H. Ban, C. Y. Kang, S. D. Hwang, S. R. Choi, et al., "Trends in epidemiologic characteristics of end-stage renal disease from 2019 Korean Renal Data System (KORDS)'', *Kidney research and clinical practice,* Vol.40, No.1, pp.52-61, 2021.
DOI: https://doi.org/10.23876/j.krcp.20.202

[8] S. Y. Kim, *A Analysis in Complication Prediction Using Machine Learning Prediction Algorithm : Focusing on National Health Data*, Master's thesis, Namseoul University, Cheonan, Korea, pp.1-4, 2019.

[9] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, J. Sun, "Doctor AI: Predicting clinical events via recurrent neural network", *Proceeding of Machine Learning for Healthcare, JMLR W&C Track,* Vol.56, pp.1-16, August 2016. Available From: http://proceedings.mlr.press/v56/Choi16.pdf

[10] J. A. O'Brien, A. R. Patrick, J. Caro, "Estimates of direct medical costs for microvascular and macrovascular complications resulting from type 2 diabetes mellitus in the United States in 2000", *Clinical therapeutics,* Vol.25, No.3, pp.1017-1038, 2003.
DOI: https://doi.org/10.1016/s0149-2918(03)80122-4

[11] Z. C. Lipton, D. C. Kale, C. Elkan, R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks", arXiv preprint [cited 2015 November 11], Available From: https://arxiv.org/pdf/1511.03677v4.pdf (accessed March. 20, 2022)

[12] B. J. Lee, "Prediction model of hypercholesterolemia using body fat mass based on machine learning", *The Journal of the Convergence on Culture Technology,* Vol.5, No.4, pp.413-420, 2019.
DOI: https://doi.org/10.17703/JCCT.2019.5.4.413

[13] Y. J. Jang, Y. S. Choy, C. M. Nam, K. T. Moon, E. C. Park, "The effect of continuity of care on the incidence of end-stage renal disease in patients with newly detected type 2 diabetic nephropathy: a retrospective cohort study", *BMC nephrology,* Vol.19, No.1, pp.1-12, 2018.
DOI: https://doi.org/10.1186/s12882-018-0932-3

[14] V. Gulshan, L. Peng, M. Coram, M. Coram, M. C. Stumpe, D. W, A. Narayanaswamy et al., "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs", JAMA. Vol. 316, No.22, pp.2402-2410, 2016. Available From: https://jamanetwork.com/journals/jama/article-abstract/2588763

[15] L. A. Stevens, J. Coresh, T. Greene, A. S. Levey, "Assessing kidney function—measured and estimated glomerular filtration rate", *New England Journal of Medicine,* Vol.354, No.23, pp.2473-2483, 2006.
DOI: https://doi.org/10.1056/NEJMra054415

[16] National Kidney Foundation [Internet]. U. S. Available From: https://www.kidney.org/atoz/content/gfr (accessed Jan. 2, 2022)

[17] The Korean socierty of nephrolgy [Internet]. Korea. Available From: https://ksn.or.kr/general/about/check.php (accessed Jan. 2, 2022)

[18] N. K. Hong, H. J. Park, Y. M. Rhee, "Machine learning application in diabetes and endocrine disorders",. *The Journal of Korean Diabetes,* Vol.21, No.3, pp.130-139, 2006.
DOI: https://doi.org/10.4093/jkd.2020.21.3.130

[19] E. S. Choi, N. J. Park, "Application and Development of Machine Learning Training Program based on Understanding K-NN Algorithm'', *Journal of The Korean Association of Information Education,* Vol.25, No.1, pp.175-184, 2021.
DOI: https://doi.org/10.14352/jkaie.2021.25.1.175

[20] J. S. Chou, C. F. Tsai, A. D. Pham, Y. H. Lu, "Machine learning in concrete strength simulations: Multi-nation data analytics", *Construction and Building materials,* Vol.73, pp.771-780, 2014.
DOI: https://doi.org/10.1016/j.conbuildmat.2014.09.054

[21] M. Zafari, D. Kumar, M. Umer, K. S. Kim, "Machine learning-based high throughput screening for nitrogen fixation on boron-doped single atom catalysts", *Journal of Materials Chemistry A*, Vol.8, No.10, pp.5209-5216, 2020.
DOI: https://doi.org/10.1039/C9TA12608B

[22] W. Li, Y. Yin, X. Quan, H. Zhang, "Gene expression value prediction based on XGBoost algorithm", *Frontiers in genetics,* Vol.10, Article 1077, 2019.
DOI: https://doi.org/10.3389/fgene.2019.01077

[23] D. H. Kim, M. K. Jwa, S. J. Lim, S. M. Park, J. W. Joo, "A Study on the Prediction Algorithm of Diabetes Based on XGBoost: Data from the 2016~2018 Korea National Health and Nutrition Examination Survey", *The Korean Institute of Communications and Information Sciences*, pp.965-966, February 2021.

[24] I. J. Lee, J. Lee, "Predictive of Osteoporosis by Tree-based Machine Learning Model in Post-menopause Woman", *Journal of radiological science and technology,* Vol.43, No.6, pp.495-502, 2020.
DOI: https://doi.org/10.17946/JRST.2020.43.6.495

[25] G. N. Dimitrakopoulos, A. G. Vrahatis, V. Plagianakos, K. Sgarbas, "Pathway analysis using XGBoost classification in Biomedical Data", *Proceedings of the 10th Hellenic Conference on Artificial Intelligence,* Patras Greece, Article No.46, pp.1-6, July 2018.
DOI: https://doi.org/10.1145/3200947.3201029

[26] T. Chen, 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, August 2016.
DOI: https://doi.org/10.1145/2939672.2939785

[27] M. A. Sarwar, N. Kamal, W. Hamid, M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare", 2018 24th international conference on automation and computing (ICAC), IEEE, Newcastle Upon Tyne, UK, September 2018.
DOI: https://doi.org/10.23919/IConAC.2018.8748992

[28] X. Chen, Q. Yin, L. Ma, P. Fu, "The Role of Cholesterol Homeostasis in Diabetic Kidney Disease", Current Medicinal Chemistry, Vol.28, pp.7413-7426, 2021.
DOI: https://doi.org/10.2174/0929867328666210419132807

[29] R. Z. Alicic, M. T. Rooney, K. R. Tuttle, "Diabetic kidney disease:challenges, progress, and possibilities", Clinical Journal of the American Society of Nephrology, Vol.12, pp.2032-2045, 2017.
DOI: https://doi.org/10.2215/CJN.11491116

[30] Q. Zhaoa, X. Yib, Z. Wanga, "Meta-Analysis of the Relationship betweenAbdominal Obesity and Diabetic Kidney Diseasein Type 2 Diabetic Patients", Obes Facts, No.14, pp.338-345, 2021.
DOI: https://doi.org/10.1159/000516391

[31] J. Jung, N. Lee, S. Kim, G. Seo, Oh, H. "Diabetes prediction mechanism using machine learning model based on patient IQR outlier and correlation coefficient.", Journal of the Korea Institute of Information and Communication Engineering, No.25, pp. 1296-1301, 2021.
DOI: http://doi.org/10.6109/jkiice.2021.25.10.1296

박 윤 진(Yoonjin Park) [정회원]

- 1997년 2월 : 국군간호사관학교간호학과 (간호학 학사)
- 2004년 8월 : 경희대학교 행정대학원 사회복지학과 (사회복지학 석사)
- 2017년 9월 : 가톨릭대학교 간호학과 (간호학 박사)
- 2019년 9월 ~ 현재 : 중부대학교 간호학과 교수

〈관심분야〉
재활간호, 호스피스 간호

강 혜 경(Hyekyung Kang) [정회원]

- 1998년 2월 : 국군간호사관학교간호학과 (간호학 학사)
- 2010년 8월 : 한양대학교 간호학 석사
- 2014년 8월 : 한양대학교 간호학 박사
- 2022년 3월 ~ 현재 : 중부대학교 간호학과 교수

〈관심분야〉
성인간호, 노인 및 치매간호