KJFP https://doi.org/10.21215/kjfp.2022.12.3.173
eISSN 2233-9116
Korean J Fam Pract. 2022;12(3):173-178

Original Article

Korean Journal of Family Practice

고혈압에 대한 기계 학습 기반 위험 요인 분석: 국민건강영양조사 2016–2019

오태섭[1], 김동균[2], 원장원[1], 김선영[1], 정은진[1], 양지수[1], 유정화[1], 김병성[1,*], 이주현[2,*]
[1]경희대학교병원 가정의학과, [2]한양대학교 전자공학과

A Machine-Learning-Based Risk Factor Analysis for Hypertension: Korea National Health and Nutrition Examination Survey 2016–2019

Taeseob Oh[1], Dongkyun Kim[2], Changwon Won[1], Sunyoung Kim[1], Eunjin Jeong[1], Jisoo Yang[1], Junghwa Yu[1], Byungsung Kim[1,*], Joohyun Lee[2,*]
[1]Department of Family Medicine, Kyung Hee University Medical Center, Seoul; [2]Department of Electrical and Electronic Engineering, Hanyang University, Ansan, Korea

Background: The purpose of this study was to use machine learning to identify risk factors (other than systolic and diastolic blood pressure) for hypertension.

Methods: The study population comprised 23,170 adults (selected from the KNHANES 2016–2019), of whom 7,500 (32.4%) had hypertension. We developed machine learning-based classification models for diagnosing hypertension using the computerized demographic and examination survey database of subjects from the KNHANES study. Random forest (RF)- and gradient boosting machine (GBM)-based classification algorithms were trained with 5-fold cross-validation, and factors related to hypertension were identified through post-hoc analysis using the permutation feature importance (PFI) technique. The classifiers used 59 variables whose data could be easily extracted on medical examination, excluding directly related variables like systolic and diastolic blood pressure.

Results: The classification performance of GBM (area under the curve [AUC], 0.852; 95% confidence interval [CI], 0.842–0.862) was slightly higher than that of RF (AUC, 0.847; 95% CI, 0.837–0.857). Post-hoc analysis of model classification using the PFI technique revealed age, cholesterol level, fraternal hypertension, education level, and height as risk factors for hypertension.

Conclusion: Although hypertension diagnosis is based on systolic and diastolic blood pressure measurements, hypertension could also be diagnosed by analyzing easily extractable variables such as age, cholesterol level, and family history of hypertension using machine learning.

Keywords: *Hypertension; Machine Learning; Risk Factors; Data Mining; Artificial Intelligence*

서 론

고혈압은 주요 만성질환으로 높은 유병률과 합병증으로 인한 사망이 매우 높은 질환이다.[1)] 고혈압은 뇌졸중 및 관상동맥 질환의 주요 위험 인자이며, 고혈압 환자는 신부전, 심부전, 말초혈관질환 등 합병증에 걸리기 쉽다.[2)] 전 세계적으로 14억 명의 성인(18세 이상 여성의 20%, 남성의 25%)이 고혈압을 앓고 있는 것으로 추정된다.[2)] 고혈압은 심뇌혈관질환의 주요 위험인자 중 하나이며, 전 세계적으로

Received August 19, 2021 **Revised** April 17, 2022
Accepted April 27, 2022
Corresponding author Byungsung Kim
Tel: +82-2-958-8700, Fax: +82-2-958-8699
E-mail: byungsungkim@naver.com
ORCID: https://orcid.org/0000-0002-3293-9640
Corresponding author Joohyun Lee
Tel: +82-31-400-5662, Fax: +82-31-436-8102
E-mail: joohyunlee@hanyang.ac.kr
ORCID: https://orcid.org/0000-0002-7698-1568

사망과 질병 부담 측면에서 커다란 문제가 되고 있다.[3]

최근 빅데이터 기반의 기계 학습 분류기를 이용하여 환자의 질병을 분류하거나 위험도를 예측하는 연구가 널리 시행되고 있다. 기계 학습을 활용한 임상적용 연구의 예시로, 국민건강영양조사 데이터를 활용하여 심근경색 및 협심증 발생을 머신 러닝으로 예측하는 연구가 수행되었으며, 국민건강보험공단에서 제공하는 공공데이터를 이용하여 뇌졸중 위험도를 높은 정확도로 예측하는 기계 학습 모델이 제시되었다.[4,5] 여기에는 제한점도 있는데 기계 학습 기반 분류기들은 높은 정확도를 보이지만, 대부분 의사 결정 과정을 설명할 수 없다는 단점이 있다. 최근에는 해석 가능한 모델인 의사결정나무로 심혈관 질환자의 재입원 여부를 분류하고, 위험요인 분석으로 모델을 해석한 연구가 수행되었다.[6]

한편, 고혈압을 대상으로 한 연구에서 Logistic regression과 Naïve bayes와 같은 모수적 모델을 기반으로 고혈압의 위험인자를 분석한 연구도 보고되고 있으나,[7] 아직 비선형적이고 비모수적인 기계 학습 기법을 기반으로 고혈압을 예측하고 모델 해석을 수행한 연구는 찾아보기 어렵다.

본 연구의 목적은 수축기 및 이완기 혈압을 제외한 정보를 이용하여 트리 앙상블 기반의 분류기들을 생성하고, 변수 중요도 기반의 모델 해석 기술인 permutation feature importance (PFI)로 고혈압의 위험인자들을 파악하는 것을 목적으로 한다.

방 법

1. 연구 대상 및 변수

본 연구는 경희대학교병원의 IRB (2020-06-054) 승인을 받아 진행하였으며 국민건강영양조사(Korea National Health And Nutrition Examination Survey, KNHANES) 2016–2019년 자료의 참가자를 대상으로 고혈압의 위험인자를 분석한 횡단 연구(cross-sectional)이다. KNHANES 2016–2019로부터 추출된 32,379명의 참가자 중 만 19세 이하의 참가자 6,384명과 고혈압에 대한 유병 여부가 결측된 참가자 2,825명을 제외하여, 최종적으로 23,170명의 데이터 세트를 구성하였다. 데이터 세트의 변수는 KNHANES가 2016년부터 2019년까지 공통으로 조사한 695개의 변수 중에 성별, 나이, 가구소득, 교육 수준, 가족력, 혈압측정, 신체 계측, 혈액검사, 소변검사, 호흡기증상만을 포함해 구성하였다. 변수들에 존재하는 결측값들은 null값으로 대치되었으며, 다시 null값은 이산형 변수는 최빈값, 수치형 변수는 중앙값으로 대치되었다. 결측값이 전체 표본의 20%를 넘어가는 변수는 제외하였고, 고혈압 진단에 직접적으로 연관된 수축기/이완기 혈압 또한 제외하여, 총 59개의 입력 변수 집합을 구성하였다. 본 연구에 사용된 변수들의 통계학적 정보들을 Table 1과 Table 2에 기술하였다.

2. 고혈압 정의

고혈압은 KNHANES의 이용지침서를 참고하여 수축기 혈압이 140 mmHg 이상이거나 이완기 혈압이 90 mmHg 이상 또는 고혈압

Table 1. Characteristics of demographic information and blood examination variables

Variable	Normal (n=15,670)	Hypertension (n=7,500)	P-value[a]
Age (y)	45.92±15.44	62.53±13.23	<0.001
Female	9,068 (57.87)	3,869 (51.59)	<0.001
Monthly Income			<0.001
Low	3,742 (23.88)	1,932 (25.76)	
Middle low	3,886 (24.80)	1,891 (25.21)	
Middle high	3,977 (25.38)	1,849 (24.65)	
High	4,018 (25.64)	1,793 (23.91)	
Education level			<0.001
≤Elementary school	1,656 (10.57)	2,648 (35.31)	
Middle school	1,177 (7.51)	1,015 (13.53)	
High school	5,288 (33.75)	1,921 (25.61)	
≥College	6,897 (44.01)	1,537 (20.49)	
Pulse frequency (min)	70.84±8.63	70.94±9.23	0.434
Irregular pulse	113 (0.72)	158 (2.11)	<0.001
Weight (kg)	63.33±12.21	65.69±12.90	<0.001
Height (cm)	164.32±8.98	161.08±9.72	<0.001
BMI (kg/m^2)	23.35±3.37	25.20±3.57	<0.001
Waist circumference (cm)	80.53±9.86	87.43±9.54	<0.001
Fasting time (h)	13.22±2.17	13.34±2.01	0.180
Blood examination			
Fasting glucose (mg/dL)	97.21±19.74	109.41±28.03	<0.001
HbA1c (%)	5.58±0.69	6.04±0.96	<0.001
Cholesterol (mg/dL)	194.58±36.24	187.87±40.25	<0.001
HDL (mg/dL)	52.77±12.71	48.90±12.23	<0.001
Triglyceride (mg/dL)	123.89±96.20	155.08±122.24	<0.001
HBsAg (IU/mL)	94.78±649.90	87.61±625.47	0.487
HBsAg positive	469 (2.99)	227 (3.03)	0.921
AST (IU/L)	22.33±11.99	25.84±17.64	<0.001
ALT (IU/L)	21.31±18.25	24.86±18.26	<0.001
Hepatitis C antibody (IU/mL)	0.12±0.73	0.15±0.90	0.002
Hemoglobin (g/dL)	14.03±1.61	14.07±1.59	0.150
Hematocrit (%)	42.77±4.38	42.75±4.46	0.515
Blood urea nitrogen (mg/dL)	14.20±4.16	16.52±5.59	<0.001
Creatinine (mg/dL)	0.80±0.20	0.86±0.34	<0.001
White blood cell (Thous/uL)	6.12±1.73	6.43±1.83	<0.001
Red blood cell (Mil/uL)	4.62±0.49	4.57±0.52	<0.001
Platelet (Thous/uL)	261.50±62.61	255.09±68.21	<0.001
Uric acid (mg/dL)	5.01±1.33	5.29±1.47	<0.001

Values are presented as mean±standard deviation or number (Proportion, %). The dataset include missing values, however we did not consider them when we calculate statistics.
BMI, body mass index; HbA1c, glycated hemoglobin; HDL, high density lipoprotein; HBsAg, hepatitis B surface antigen; AST, aspartate aminotransferase; ALT, alanine aminotransferase.
[a]P-values were derived from χ^2 test or Mann–Whitney U test.

Table 2. Characteristics of urine examination and family history

Variable	Normal (n=15,670)	Hypertension (n=7,500)	P-value[a]
Urine test			
Urine pH	5.81±0.78	5.88±0.84	<0.001
Urine nitrite	286 (1.83)	269 (3.59)	<0.001
Urine specific gravity	1.020±0.007	1.018±0.006	<0.001
Urine creatinine (mg/dL)	158.89±86.66	129.87±73.44	<0.001
Urine natrium (mmol/L)	112.00±48.72	115.52±46.72	<0.001
Urine kalium (mmol/L)	50.64±23.56	51.85±23.29	0.002
Dipstick test[b]			
Urine protein	0.16±0.43	0.26±0.64	<0.001
Urine glucose	0.08±0.53	0.23±0.88	<0.001
Urine ketone	0.08±0.41	0.05±0.27	<0.001
Urine bilirubin	0.09±0.42	0.07±0.37	<0.001
Urine blood	0.42±0.95	0.42±0.86	<0.001
Urine urobilinogen	0.02±0.16	0.01±0.16	0.194
Respiratory symptom			
Cough	232 (1.48)	230 (3.07)	<0.001
Sputum	364 (2.32)	345 (4.60)	<0.001
Family medical history			
Chronic disease[c]	9,269 (59.15)	4,675 (62.33)	<0.001
Paternal diabetes	1,676 (10.70)	449 (5.99)	<0.001
Maternal diabetes	1,726 (11.01)	729 (9.72)	0.395
Fraternal diabetes	933 (5.95)	794 (10.59)	<0.001
Paternal hyperlipidemia	412 (2.63)	94 (1.25)	<0.001
Maternal hyperlipidemia	756 (4.82)	195 (2.60)	<0.001
Fraternal hyperlipidemia	277 (1.77)	152 (2.03)	0.048
Paternal stroke	835 (5.33)	608 (8.11)	<0.001
Maternal stroke	593 (3.78)	565 (7.53)	<0.001
Fraternal stroke	176 (1.12)	241 (3.21)	<0.001
Paternal hepatitis	134 (0.86)	22 (0.29)	<0.001
Maternal hepatitis	149 (0.95)	34 (0.45)	<0.001
Fraternal hepatitis	176 (1.12)	60 (0.8)	0.076
Paternal hypertension	2,927 (18.68)	1,351 (18.01)	0.014
Maternal hypertension	3,338 (21.30)	1,978 (26.37)	<0.001
Fraternal hypertension	1,077 (6.87)	1,716 (22.88)	<0.001

Values are represented as mean±standard deviation or number (%).
[a]P-values were derived from χ^2 test or Mann–Whitney U test. [b]0: -, 1: ±, 2: +, 3: ++, 4: +++, 5: ++++. [c]Chronic disease: Hypertension, Hyperlipidemia, Ischemic heart disease, stroke, diabetes mellitus, thyroid disease, Hepatitis B.

치료제를 복용 중이면 고혈압 환자군으로 정의하였고, 그 외 참가자들은 모두 정상군으로 정의하였다.[8]

3. 분류 모델 훈련

KNHANES 2016–2018년 데이터로는 학습 셋(train set)을 구성하였고. KNHANES 2019년 데이터로는 시험 셋(test set)을 구성하였다. 기계 학습 분류기는 train set의 입력 변수와 고혈압 여부를 동시에 입력 받아 학습하는 지도 학습 방법론으로 훈련하였다. 추론 시에는 test set을 대상으로 질병 유무를 예측하고 참값과 비교하여 성능을 측정했다.

분류 모델로는 다수의 의사결정 트리를 학습하여 예측하는 트리-앙상블 모델인 random forest (RF)와 gradient boosting machine (GBM)을 사용하였다. 트리 기반 모델들은 데이터에 대한 가정이 없어 자유롭게 사용할 수 있고, 데이터 전처리가 따로 필요하지 않다는 장점이 있다. 각 트리들은 이전 트리의 분류 오차를 경사 하강법으로 보완하여 좀 더 정확한 예측을 하도록 훈련된다. 이러한 훈련 방식은 모델이 강한 분류 능력을 갖추게 하여 예측의 편향을 줄인다. 따라서 복잡한 고차원의 데이터를 분류하는 데 용이하다.[9]

기계 학습 분류기들이 최상의 성능을 내기 위해서는 데이터 세트에 적합한 초 매개변수(hyperparameter)를 탐색해야 한다. 이를 위해서 본 연구에서는 5-fold 교차 검증과 grid-탐색 방법을 결합하여 사용하였다. 5-fold 교차 검증은 train set을 고혈압 환자의 비율을 같이 유지하면서 5등분한 뒤, 그 중 4개의 데이터 세트로 모델을 훈련하고 나머지 1개의 데이터로 모델 성능을 측정한다. 5개의 등분에 대해 훈련과 측정을 번갈아 실행하여 얻은 성능의 평균을 이용하는 방법론이다.[10] 본 연구에서는 5-fold 교차 검증의 성능 지표로, area under the receiver operating characteristics curve (AUC)를 사용하였다. grid-탐색은 미리 설정된 초 매개변수의 모든 경우의 수에서 교차 검증의 성능이 가장 높은 초 매개변수 조합을 찾는 방법론이다.

본 연구에서 사용된 데이터의 고혈압 환자군과 정상군의 비율은 32:68로 불균형하다(Table 1). 일반적으로 불균형한 데이터를 기반으로 학습한 모델은 다수를 차지하는 부류에 과적합되어 소수 부류에 대해 낮은 분류 성능을 보이게 된다. 이러한 문제는 모델 훈련 시, 샘플링 기법을 통해 부류별 표본 수를 균등하게 만들어 줌으로써 해결할 수 있다. 본 연구에서는 random over sampling 기법으로 소수 부류에 해당하는 유병자들을 무작위로 복제하여 정상 표본들의 수와 같아지도록 했다. 해당 샘플링 기법은 5-fold 교차 검증과정에서 훈련용으로 사용된 4개의 fold에 적용되었다.

4. 통계분석

본 연구의 입력변수와 고혈압 간의 통계학적 특성을 분석하기 위해서, 데이터 세트를 고혈압군과 정상군으로 나눈 뒤 변수들을 양성 응답 수와 비율 또는 평균과 표준편차로 정리하여 Table 1과 Table 2에 기술하였다. 학습이 완료된 분류 모델들의 성능은 test set에서의 정확도(accuracy), 특이도(specificity), 민감도(sensitivity), geometric mean (G-mean), AUC로 비교하였다. 그런데 일반적으로 데이터가 음성 부류에 치우쳐져 있으면, 민감도는 감소하고 특이도는 증가하게 된다. 따라서 본 연구에서는 추가적으로 민감도와 특이도의 기하 평균인 G-mean을 사용하여 소수 부류와 다수 부류에서의 성능을 균형적으로 평가했다.[11] 모든 성능 지표는 bootstrapping을 이용한 95%

신뢰 구간(95% confidence interval [CI])과 평균으로 나타냈다.[12] 본 연구의 모델 생성 및 통계분석을 위해 Python 3.8 (Python Software Foundation, Wilmington, DE, USA) 언어와 Scikit-learn 0.24[13], SciPy 1.6[14] 패키지를 이용하였다.

5. 변수 중요도

다음으로 학습이 완료된 고혈압 분류 모델의 중요 위험인자를 PFI를 이용하여 파악했다. PFI는 각 변수가 모델의 의사결정에 공헌한 정도를 상대적으로 정량화하는 알고리즘이다. PFI는 공헌도를 측정하고자 하는 변수를 무작위로 섞었을 때, 모델의 성능이 감소하는 수치를 계산하여 변수의 중요도를 측정한다. 본 연구에서는 모든 변수별로 PFI 알고리즘을 100회 시행하여 감소한 AUC의 평균으로 위험인자의 중요도 순위를 결정하였다.

$$\mathrm{PFI}=\mathrm{AUC}(N)-\mathrm{AUC}(N\backslash\{i\}\cup\{i'\})$$

N은 모든 변수를 포함한 데이터 집합, i는 중요도 측정 대상이 되는 변수, i'는 해당 변수를 무작위로 섞은 것이다.

결 과

KNHANES 2016–2019로부터 추출한 연구 모집단은 총 23,170명이었으며, 그 중에서 고혈압 환자는 7,500명, 정상군은 15,670명이었다(Table 1). 나이는 고혈압 환자군(평균 62.53세)이 정상군(평균 45.92세)보다 높았고, 여성의 비율은 고혈압 환자군(51.59%)이 정상군(57.87%)보다 낮았다. 불규칙 혈압(irregular pulse)의 비율, 몸무게, 허리둘레 또한 고혈압 환자군이 높았다. 혈액 검사 변수에서는 고혈압 환자군의 공복혈당(fasting glucose), 당화혈색소(HbA1c), 중성지방(triglyceride), aspartate aminotransferase (AST), alanine aminotransferase (ALT), 혈중요소질소(blood urea nitrogen)가 두드러지게 높았다.

Table 2에 기술된, 소변 검사에서는 고혈압 환자군의 요 크레아티닌(urine creatinine)의 검진 수치가 정상군보다 낮았으며, 가족력 변수의 경우, 형제자매에 대한 당뇨(fraternal diabetes) 또는 고혈압(fraternal hypertension)에 대한 양성 응답 비율이 환자군에서 더 높았다.

본 연구에서는 앞서 설명한 변수들을 이용하여 고혈압 분류를 수행하는 트리 앙상블모델을 생성하고, test set에서의 분류 성능을 비교하였다. Table 3은 59개 입력 변수를 이용하여 훈련한 RF와 GBM의 분류 성능을 비교한 것이다. 정확도는 RF (accuracy, 0.761; 95% CI, 0.750–0.771)가 GBM (accuracy, 0.759; 95% CI, 0.749–0.769)보다 높았고, 특이도 또한 RF 모델(specificity, 0.740; 95% CI, 0.726–0.752)이 GBM (specificity, 0.732; 95% CI, 0.718–0.745)보다 소폭 높은 것으로 보아, RF가 GBM보다 정상군을 더 잘 분류해내는 것을 알 수 있다. 민감도는 GBM 모델(sensitivity, 0.817; 95% CI, 0.799–0.834)이 소폭 더 높으므로, GBM 모델이 고혈압 환자를 좀 더 잘 분류하는 것으로 여겨진다. 균형적인 성능 지표인 G-mean 및 AUC의 결과에 따르면, GBM 모델(G-mean 0.773, 95% CI 0.762–0.783; AUC 0.852, 95% CI 0.842–0.862)이 RF 모델(G-mean 0.771, 95% CI 0.761–0.782; AUC 0.847, 95% CI 0.837–0.857)보다 분류 성능이 높았다.

Table 4는 고혈압 분류에 대한 상대적인 중요도를 PFI로 기술한 것으로, 모델별 고혈압 분류에 공헌이 가장 컸던 상위 10개 변수를 나타내고 있다. 두 모델에서 공통으로 나이가 가장 중요한 변수이면서, 다른 변수들과의 중요도 차이가 매우 컸다. 또한 공통으로 콜레스테롤 수치, 형제자매의 고혈압 여부, 참가자의 교육 수준, 신장, 크레아티닌 수치가 상위에 배치되었다. 두 모델의 차이점으로는, RF에서는 요산도, 혈당, 요 비중이, GBM에서는 요 단백, 요 나트륨, 적혈구 수, 백혈구 수가 상위 10개 변수에 포함되었다.

Table 3. Classification performance of random forest and gradient boosting machine

Metrics	Random forest (95% CI)	Gradient boosting machine (95% CI)
Accuracy	0.761 (0.750–0.771)	0.759 (0.749–0.769)
Specificity	0.740 (0.726–0.752)	0.732 (0.718–0.745)
Sensitivity	0.805 (0.786–0.822)	0.817 (0.799–0.834)
G-mean	0.771 (0.761–0.782)	0.773 (0.762–0.783)
AUC	0.847 (0.837–0.857)	0.852 (0.842–0.862)

G-mean, geometric mean; AUC, area under the receiver operating characteristics curve.

Table 4. Permutation feature importance of hypertension classification

Random forest	Gradient boosting model
Age (0.059)	Age (0.085)
Cholesterol (0.007)	Cholesterol (0.011)
Fraternal hypertension (0.006)	Education level (0.007)
Education level (0.004)	Fraternal hypertension (0.006)
Urine acid (0.003)	Urine protein (0.004)
Height (0.003)	Urine natrium (0.004)
Creatinine (0.003)	Height (0.004)
Maternal hypertension (0.002)	Creatine (0.003)
Glucose (0.002)	Red blood cell count (0.003)
Urine specific gravity (0.001)	White blood cell count (0.003)

고 찰

본 연구는 빅데이터에서 추출한 참가자의 기본 변수, 검진 조사 변수들을 통합하고, 혈압에 대한 변수를 제외한 상태에서 고혈압 유병 여부를 분류하는 기계 학습 모델을 제시하였다. 제시된 RF 모델과 GBM 모델은 수축기 및 이완기 혈압 변수가 제외되었음에도 0.8 이상의 AUC 성능을 보였다. 혈압 이외에 고혈압의 위험 인자로 나이, 콜레스테롤, 고혈압 가족력, 교육 수준 등이 있음을 밝혔다. 또한 본 연구는 PFI가 높은 고혈압의 상위 10개 위험인자를 분석하였으며 이 중 나이가 고혈압 발병의 가장 주요 위험인자임을 확인하였다. 특히 나이가 들면 혈관의 탄력이 떨어지면서 혈압이 서서히 높아지는 경향이 있기 때문에 고혈압의 위험은 나이가 들수록 높아진다. 또한 가족 구성원은 자신의 건강과 질병 위험에 영향을 미칠 수 있는 유전자, 행동, 생활 방식 및 환경을 공유하여 가족력이 있을 수 있으며 고혈압의 위험은 연령과 인종 또는 민족에 따라 증가할 수 있다.[15]

이 연구는 우리나라 인구에서 다양한 기계 학습 모델을 이용하여 고혈압 위험인자의 중요성을 평가한 첫 번째 연구이며 이전 연구 결과와 일치하였다. 기존 연구결과에 따르면 젊은 성인에서 10년 동안 발생한 고혈압은 연령과 초기 수축기 혈압에 의해 가장 강력하게 예측되었으며, 그 외 BMI, 허리둘레, 신체활동, 음주, 맥박수, 흡연, 교육수준, 당뇨, 이상지질혈증, 요산 등이 영향을 미쳤다.[16] AlKaabi 등[17]의 연구에서는 사회인구학적 정보뿐만 아니라, 과일 및 채소의 영양분 섭취, 부모의 과거병력 등에 대한 정보를 바탕으로 RF, Decision Tree, Logistic Regression 알고리즘들을 이용하여 고혈압 예측 모델을 제시하였고, 0.799–0.869의 AUC값을 획득하였다. Zhao 등[18]의 연구에서는 18–70세의 대상자에 대해 RF, CatBoost, Multi-layer Perceptron neural network, Logistic Regression 모델을 이용 10개의 선택된 변수(연령, 성별, BMI, 허리둘레, 가족력, 직업, 흡연, 음주, 식단, 신체활동)를 훈련과 검증을 통해 RF가 0.92의 AUC, 0.82의 정확도, 0.83의 민감도, 0.81의 특이도일 때 가장 좋은 성능을 보인다고 결론지었다.

주요 위험인자를 파악하는 것은 고혈압 위험평가에서 임상적으로 매우 중요하다. 고혈압은 한 가지 원인에 의해 유발되지 않고 여러 가지 요인이 모여 고혈압을 일으킨다. 고혈압의 다양한 원인을 고려하기 위해 국민건강영양조사에서 설문조사를 제외한 모든 데이터를 변수로 설정하여 기계 학습시켰다. 이는 기존에 연구되지 않은 위험인자를 찾고자 하는 연구목적에서 이루어졌다. 본 연구는 또한 다양한 위험요인의 중요성이 모델링 기술에 달려있음을 보여주었다. RF에서는 요산도, 혈당, 요 비중이, GBM에서는 요 단백, 요 나트륨, 적혈구 수, 백혈구 수가 고혈압의 상위 10개 위험인자에 포함되었다.

본 연구의 제한점으로는 첫째, 국민건강영양조사는 추적조사 없이 데이터를 분석한 단면연구이기 때문에 고혈압의 위험인자와 인과관계의 방향을 설명하는데 한계가 있다. 둘째, 모델의 성능을 평가하기 위해서는 향후 외부 검증 및 다른 기계 학습 모델을 통한 연구가 이루어져야 할 것이다. 셋째, 다양한 목적을 위해 수행된 국민건강영양조사를 연구하였기 때문에 고혈압 위험인자의 모든 변수가 포함되지 못하였으며 데이터의 일부만 사용했기 때문에 고혈압과의 연관성에 대해 충분히 의미 있게 분석하지 못한 한계가 있다. 넷째, 국민건강영양조사의 설문조사 시 응답자의 상황에 따라 일부 편견이 들어갈 수 있고 개인의 기억에 의존하여 작성되어 응답자의 회상 오류가 발생할 수 있다. 다섯째, 고혈압 치료를 이미 시작한 사람들의 영향을 배제할 수 없었다. 앞으로의 연구에서는 이러한 고혈압 환자를 제외함으로써 치료효과를 배제할 수 있다.

이러한 제한점에도 불구하고, 본 연구는 우리나라 인구를 대표할 수 있는 대규모 인구집단을 대상으로 한 국민건강영양조사 자료를 이용하여 고혈압에 영향을 미칠 수 있는 다양한 요인을 분석하였다는 점에서 의의가 있다고 여겨진다.

결론적으로, 고혈압에서 나이, 콜레스테롤, 고혈압 가족력, 교육 수준이 주요 위험인자로 작용하는 것을 확인하였다. 본 연구결과를 통해 이전 위험요인 평가 연구에서 발견되지 않은 새로운 위험요인(요산도, 요 비중, 요 단백, 요 나트륨)이 발견되었다. 기계 학습은 고혈압의 위험을 예측하는 능력을 보여 주었고, 확진을 위한 진단검사 이전에 간단한 병력청취와 혈액, 소변검사로 고혈압의 발생 가능성을 예측해 볼 수 있다. 이러한 결과는 우리나라 고혈압 발생 고위험군을 선별하거나 고혈압으로 인한 합병증을 예방할 수 있는 중재 프로그램 개발에 유용한 근거자료가 될 것이다.

요 약

연구배경: 본 연구의 목적은 수축기 및 이완기 혈압을 제외한 정보를 이용하여 트리 앙상블 기반의 분류기들을 생성하고, 변수 중요도 기반의 모델 해석 기술인 permutation feature importance (PFI)로 고혈압의 위험인자들을 파악하는 것이다.

방법: 연구 인구는 23,170명의 성인(국민경강영양조사 2016–2019년 자료)으로 구성되었으며 이 중 7,500명(32.4%)이 고혈압이 있었다. 국민경강영양조사의 자료를 사용하여 고혈압 진단을 위한 기계 학습 기반 분류 모델을 개발했다. Random forest (RF)와 gradient boosting machine (GBM) 기반의 분류 알고리즘은 5-fold 교차 검증으로 학습되었고, PFI 기법을 이용한 사후분석을 통해 고혈압과 관련된 요인

을 식별했다. 수축기 혈압과 이완기 혈압과 같이 직접적으로 관련된 변수를 제외하고 건강 진단에서 쉽게 데이터를 추출할 수 있는 59개의 변수를 사용했다.

결과: GBM의 분류 성능(area under the curve [AUC], 0.852; 95% confidence interval [CI], 0.842–0.862)은 RF (AUC, 0.847; 95% CI, 0.837–0.857)보다 약간 더 높았다. PFI 기법을 이용한 모델의 분석에서는 고혈압의 위험인자로 나이, 콜레스테롤 수치, 형제자매의 고혈압 여부, 참가자의 교육 수준, 신장이 상위에 배치되었다.

결론: 고혈압에서 나이, 콜레스테롤, 고혈압 가족력, 교육 수준이 주요 위험인자로 작용하는 것을 확인하였다. 기계 학습은 고혈압의 위험을 예측하는 능력을 보여 주었고, 확진을 위한 진단검사 이전에 간단한 병력청취와 혈액, 소변검사로 고혈압의 발생 가능성을 예측해 볼 수 있다.

중심단어: 고혈압; 기계 학습; 위험 요인; 데이터 수집; 인공지능

감사의 글

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2020-0-01343, 인공지능융합연구센터지원 한양대학교 ERICA).

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

ORCID

Taeseob Oh, https://orcid.org/0000-0001-7050-6064
Dongkyun Kim, https://orcid.org/0000-0002-6355-0101
Changwon Won, https://orcid.org/0000-0002-6429-4461
Sunyoung Kim, https://orcid.org/0000-0003-4115-4455
Eunjin Jeong, https://orcid.org/0000-0001-9166-6390
Jisoo Yang, https://orcid.org/0000-0002-3063-0316
Junghwa Yu, https://orcid.org/0000-0003-2805-0168
Byungsung Kim, https://orcid.org/0000-0002-3293-9640
Joohyun Lee, https://orcid.org/0000-0002-7698-1568

REFERENCES

1. Samadian F, Dalili N, Jamalian A. Lifestyle modifications to prevent and control hypertension. Iran J Kidney Dis 2016; 10: 237-63.
2. Mills KT, Bundy JD, Kelly TN, Reed JE, Kearney PM, Reynolds K, et al. Global disparities of hypertension prevalence and control: a systematic analysis of population-based studies from 90 countries. Circulation 2016; 134: 441-50.
3. Clement DL. Control of hypertension: is the goal reached? Hypertension 2015; 65: 25-6.
4. Lim HK. Prediction of myocardial infarction/angina and selection of major risk factors using machine learning. J Korean Data Anal Soc 2018; 20: 647-56.
5. Jeong SW, Lee MJ, Yoo SY. Machine learning-based stroke risk prediction using public big data. J Adv Navig Technol 2021; 25: 96-101.
6. Kim HS. Convergence analysis of risk factors for readmission in cardiovascular disease: a machine learning approach. J Converg Inf Technol 2019; 9: 115-23.
7. Heo BM, Kim SY, Ryu KH. Statistical analysis for risk factors and prediction of hypertension based on health behavior information. J Digit Contents Soc 2018; 19: 685-92.
8. Korea Centers for Disease Control and Prevention. The seventh Korea National Health and Nutrition Examination Survey (KNHANES VII-3) [Internet]. Cheongju: Korea Centers for Disease Control and Prevention; 2018 [cited 2021 Jul 14]. Available from: https://knhanes.kdca.go.kr/.
9. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001; 29: 1189-232.
10. Hastie T, Friedman J, Tibshirani R. The elements of statistical learning. 2nd ed. New York (NY): Springer; 2001.
11. Kubat M, Matwin S, editors. Addressing the curse of imbalanced training sets: one-sided selection. Proceedings of the 14th International Conference on Machine Learning; 1997 Jul 8-12; Nashvilee, USA. Massachusetts: Morgan Kaufmann; 1997.
12. DiCiccio TJ, Efron B. Bootstrap confidence intervals. Stat Sci 1996; 11: 189-228.
13. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research 2011; 12: 2825-30.
14. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods 2020;17:261-72.
15. Vasan RS, Beiser A, Seshadri S, Larson MG, Kannel WB, D'Agostino RB, et al. Residual lifetime risk for developing hypertension in middle-aged women and men: The Framingham Heart Study. JAMA 2002; 287: 1003-10.
16. Dyer AR, Liu K, Walsh M, Kiefe C, Jacobs DR Jr, Bild DE. Ten-year incidence of elevated blood pressure and its predictors: the CARDIA study. Coronary Artery Risk Development in (Young) Adults. J Hum Hypertens 1999; 13: 13-21.
17. AlKaabi LA, Ahmed LS, Al Attiyah MF, Abdel-Rahman ME. Predicting hypertension using machine learning: findings from Qatar Biobank Study. PLoS One 2020; 15: e0240370.
18. Zhao H, Zhang X, Xu Y, Gao L, Ma Z, Sun Y, et al. Predicting the risk of hypertension based on several easy-to-collect risk factors: a machine learning method. Front Public Health 2021; 9: 619429.