KIPS Trans. Softw. and Data Eng.
Vol.10, No.11 pp.541~546
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2021.10.11.541

Prediction Model of Hypertension Using Sociodemographic Characteristics Based on Machine Learning

Bum Ju Lee†

ABSTRACT

Recently, there is a trend of developing various identification and prediction models for hypertension using clinical information based on artificial intelligence and machine learning around the world. However, most previous studies on identification or prediction models of hypertension lack the consideration of the ideas of non-invasive and cost-effective variables, race, region, and countries. Therefore, the objective of this study is to present hypertension prediction model that is easily understood using only general and simple sociodemographic variables. Data used in this study was based on the Korea National Health and Nutrition Examination Survey (2018). In men, the model using the naive Bayes with the wrapper-based feature subset selection method showed the highest predictive performance (ROC = 0.790, kappa = 0.396). In women, the model using the naive Bayes with correlation-based feature subset selection method showed the strongest predictive performance (ROC = 0.850, kappa = 0.495). We found that the predictive performance of hypertension based on only sociodemographic variables was higher in women than in men. We think that our models based on machine leaning may be readily used in the field of public health and epidemiology in the future because of the use of simple sociodemographic characteristics.

Keywords : Machine Leaning, Prediction Model, Hypertension, Sociodemographic Characteristics, Public Health

머신러닝 기반 사회인구학적 특징을 이용한 고혈압 예측모델

이 범 주†

요 약

최근 전 세계적으로 인공지능과 머신러닝을 기반으로 임상정보를 활용한 다양한 고혈압 식별 및 예측 모델이 개발되고 있다. 그러나 고혈압 관련 모델에 대한 대부분의 선행연구는 침습적 및 고가의 분석비용을 통한 변수들이 대부분 사용되었고, 인종과 국가의 특징에 대한 고려가 충분히 제시되지 않았다. 따라서 이 연구의 목적은 일반적인 사회인구 통계학적 변수만을 사용하여 쉽게 이해할 수 있는 한국인 성인 고혈압 예측 모델을 제시하는 것이다. 이 연구에서 사용된 데이터는 질병관리청 국민건강영양조사 (2018년)를 이용하였다. 남성에서, wrapper-based feature subset selection 메소드와 naive Bayes를 이용한 모델이 가장 높은 예측 성능 (ROC = 0.790, kappa = 0.396)을 보였다. 여성의 경우, correlation-based feature subset selection 메소드와 naive Bayes를 사용한 모델이 가장 높은 예측 성능(ROC = 0.850, kappa = 0.495)을 나타내었다. 또한 모든 모델들에서 사회인구 통계학적 변수들만을 이용한 고혈압의 예측 성능이 남성보다 여성에게서 더 높게 나타나는 것을 발견하였다. 본 연구의 결과인 machine learning 기반 고혈압 예측 모델은 한국인에 대한 단순한 사회인구학적 특성만을 사용하였기 때문에 향후 공중 보건 및 역학 분야에서 쉽게 사용될 수 있을 것으로 예상된다.

키워드 : 머신러닝, 예측모델, 고혈압, 사회인구학적 특징, 대중보건

1. 서 론

최근 의학 분야에서 다학제 연구를 기반으로 다양한 질병들을 극복하기 위한 연구들이 진행되고 있다. 이러한 최신 트렌드에 따라 인공지능 및 머신러닝 등의 분야에서도 다양한 질병들의 식별 또는 예측 모델을 개발하기 위한 노력들이 진행되어져 왔다[1-7].

고혈압은 전 세계적으로 가장 유병률이 높은 질환중 하나이다. 예를 들어, 전체 고혈압 환자 중에서 고혈압 조절 및 치료를 받는 환자 수는 미국에서 약 43%에 해당하지만 중국에서는 약 13%, 일본에서는 1/4 정도에 해당한다고 보고되었고[8], 아울러 아직까지도 자신의 고혈압 여부에 대하여 알지 못하거나, 진단/치료 받지 못하는 인구비율이 매우 높은 편으로 알려져 있다. 이러한 경향에 따라, 고혈압 진단 및 예측을 위하여 전 세계적으로 다양한 임상정보들을 바탕으로 인공지능, 딥러닝, 머신러닝, 데이터마이닝, 통계분석 등을 기반으로 고혈압에 대한 예측 모델을 개발하여 자동화된 의사결정을 지원하는 추세일 뿐만 아니라, 원격의료 등에서 고혈압 관리에 대한 새로운 전략을 수립할 수 있는 방안으로 제시되고 있다[8].

※ 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2021-0-00104, 비대면 심혈관 건강관리를 위한 디지털헬스 서비스 플랫폼 개발).
† 정 회 원 : 한국한의학연구원 디지털임상연구부 책임연구원
Manuscript Received : August 27, 2021
Accepted : October 9, 2021
* Corresponding Author : Bum Ju Lee(bjlee@kiom.re.kr)

머신러닝 등을 이용한 고혈압 식별 및 예측 모델 연구들에서 주로 사용되는 파라미터 (위험요인)로는 waist circumference, body mass index, waist-to-height ratio등과 같은 인체계측정보[9], Dual X-ray absorptiometry (DEXA)로 측정된 body composition 정보[10], 그리고 triglyceride, HDL cholesterol, LDL cholesterol 등과 같은 혈액정보와 더불어 나이, 성별, 직업, 교육, 음주, 수입 등과 같은 인구학적 정보[11] 등이 사용되고 있다.

그러나 이러한 파라미터들을 이용한 고혈압 예측 모델 또는 고혈압과의 연관성 연구의 결과는 나이, 성별, 인종, 국가, 지역 등에 따라 다르게 나타나는 경향이 있다[9,10]. 뿐만 아니라 이러한 고혈압 예측 모델에 대하여 많은 선행연구들이 수행되었음에도 불구하고 대부분의 연구들에서 제시하는 모델은 침습적 방식, 고비용, 변수 측정에서의 시간소모적인 변수 추출, 또는 일반인이 이해하기 어려운 복잡하고 전문적인 변수들의 사용에 기인하여 실제 일반인이 이해하기 어려울 뿐만 아니라, 대중보건 또는 역학 분야에서 심플하게 활용하기 어려운 모델들이 대부분이다.

따라서 본 연구의 목적은 매우 일반적이고 심플한 사회인구학적 변수들만을 사용하여 일반인이 이해하기 쉽고, 활용성이 매우 높은 고혈압 예측 모델을 제시하고자 한다. 이러한 연구의 결과인 머신러닝 기반 고혈압 예측모델은 향후 대중보건 및 역학분야에서 전문가뿐만 아니라 일반인도 쉽게 접근할 수 있을 것으로 예상한다.

2. 메소드

2.1 데이터 셋

본 연구는 질병관리청 국민건강영양조사 데이터에서 최근에 발표된 제 7기(2018년) 데이터를 이용하였다. 2018년도 데이터는 총 7992개 샘플로 구성되었으며, 이 샘플들에 대하여 다음과 같은 샘플 선별기준을 적용하였다. 1) 30-80대 나이 이외의 샘플들은 고혈압 샘플이 매우 적어 제외되었고, 2) income, education level, occupation, drinking, stress, smoking, height, weight, waist circumference, systolic blood pressure (BP), diastolic BP, AST_weekdays (주중 평균수면시간), AST_weekends (주말 평균수면시간)의 변수 정보가 없는 샘플들을 제외하였다. 최종적으로 고혈압 예측모델에 사용된 샘플수는 정상군 샘플 3672개 (남성 1555, 여성 2117), 고혈압 샘플 1423개 (남성 641, 여성 782)가 사용되었다.

고혈압 샘플의 정의는 의사로부터 고혈압 진단을 받은 피험자는 고혈압 군으로, 고혈압 진단을 받지 않은 피험자는 정상군으로 분류하였다.

본 데이터는 질병관리청 연구윤리심의위원회 승인을 받아 수집된 데이터이며(승인번호 2018-01-03-P-A), 이러한 데이터를 이용한 본 연구는 한국한의학연구원 생명윤리심의위원회로부터 승인을 획득하였다(IRB No. I-2109/008-001).

2.2 사회인구학적 정보 및 모델생성에 사용된 변수

본 연구에 사용된 사회인구학적 변수들은 크게 세 가지로 분류될 수 있다. 첫째, 인체계측정보로써 height, weight, waist circumference, body mass index, systolic BP, diastolic BP가 이에 해당한다. 두 번째로, 기본적인 사회 인구학적 정보인 age, income, education level, occupation, household number, house type, marital status가 사용되었으며, 마지막으로 고혈압의 위험요인으로 잘 알려진 drinking, stress, smoking 등이 사용되었다. 이러한 모든 사회인구학적 변수들에 대한 남녀 각각에서의 차이는 Table 1과 2에 제시하였다.

2.3 예측 모델 생성 및 통계분석 방법

본 연구를 위한 통계분석 및 모델생성은 IBM SPSS (version 23, SPSS Inc., Chicago, IL, US)와 Weka (version 3-8-3)에서 수행되었다. 모델 생성 알고리즘은 다양한 사전실험을 통하여 예측력의 우수성, 민감도 및 특이도의 타당성, 그리고 대중보건 및 역학 분야에서의 활용성 분석을 바탕으로 naïve Bayes classification 알고리즘을 선택하였다.

모델 생성에서 최적의 변수조합을 도출하기 위한 feature subset selection은 correlation-based feature subset (CFS) 메소드와 wrapper-based feature subset (wrapper) 메소드를 적용하였다. 이러한 사유로는 첫째, 변수들 간의 다중공선성 (multicollinearity) 문제를 해결하기 위하여 CFS (best first search) 메소드를 이용하였고, 둘째로는 naïve Bayes 알고리즘에 알맞은 feature subset 도출을 위하여 블랙박스 방식인 wrapper (best first search) 메소드를 선택하였다. 아울러, 전체 변수들을 활용하였을 경우와의 예측력 비교를 위하여 변수 full set을 이용한 모델 또한 제시하였다.

생성된 모델의 validation을 위해, 본 모델에 사용된 샘플수를 고려하여 5-fold cross-validation을 적용하였다. 예측력 평가를 위한 주요 성능지표로는 receiver operating characteristic curve (ROC)와 kappa 값을 이용하였으며, 구체적인 모델성능 평가를 위하여 sensitivity, 1-specificity, F-measure, Mathew correlation coefficient (MCC) 값을 제시하였다.

본 모델에 사용된 변수들에 대한 통계적 유의성 분석을 위하여, 연속형 변수들에 대한 정상군/고혈압군 샘플 간의 차이 분석에서는 독립표본 t-test (independent samples *t*-test)가 사용되었고, 범주형 변수들에 대한 분석에서는 *Chi*-square test가 사용되었다. 이러한 분석의 결과로는 p-value값을 제시하였다.

3. 실험 결과

3.1 모델링에 적용된 변수들의 통계적 유의성 분석 결과

Table 1은 남성의 고혈압 예측 모델에 사용된 사회인구학적 변수들에 대한 통계적 분석결과를 나타낸다. Table 1에

Table 1. Sociodemographic Features and Statistical Analysis in Men

Feature	Normal (n=1555)	Hypertension (n=641)	p-value
Age	51.26 (13.34)	64.53 (11.35)	<0.0001
Systolic BP	119.4 (14.81)	127.6 (15.65)	<0.0001
Diastolic BP	78.85 (10.17)	76.44 (11.12)	<0.0001
AST_Weekdays	418.6 (73.03)	424.4 (80.4)	0.1134
AST_Weekends	456.1 (84.44)	442 (91.42)	0.0006
Height	171.2 (6.653)	168 (6.317)	<0.0001
Weight	71.62 (11.45)	71.18 (10.93)	0.4093
Waist circumference	86.27 (8.5)	89.97 (8.602)	<0.0001
Body mass index	24.39 (3.213)	25.16 (3.212)	<0.0001
Income (quintile)			0.6665
Low	365 (16.6%)	166 (7.6%)	
Lower-middle	401 (18.3%)	158 (7.2%)	
Middle	398 (18.1%)	163 (7.4%)	
Upper-middle	391 (17.8%)	154 (7%)	
Education			<0.0001
Elementary school or less	172 (7.8%)	161 (7.3%)	
Middle school	133 (6.1%)	109 (5%)	
High school	481 (21.9%)	218 (9.9%)	
University or higher	769 (35%)	153 (7%)	
Occupation			<0.0001
Professionals and expert	315 (14.3%)	76 (3.5%)	
Clerks	237 (10.8%)	41 (1.9%)	
Service and sale workers	177 (8.1%)	45 (2%)	
Agricultural and fishery workers	83 (3.8%)	60 (2.7%)	
Device and assembly workers	363 (16.5%)	109 (5%)	
Elementary occupations	104 (4.7%)	61 (2.8%)	
Unemployed	276 (12.6%)	249 (11.3%)	
Household members			<0.0001
1	149 (6.8%)	88 (4%)	
2	425 (19.4%)	305 (13.9%)	
3	413 (18.8%)	122 (5.6%)	
4	410 (18.7%)	94 (4.3%)	
5	125 (5.7%)	23 (1%)	
6 and more	33 (1.5%)	9 (0.4%)	
House type			<0.0001
Detached house	453 (20.6%)	253 (11.5%)	
Apartment	921 (41.9%)	296 (13.5%)	
Multifamily house	166 (7.6%)	89 (4.1%)	
Etc.	15 (0.7%)	3 (0.1%)	
Marital status			<0.0001
Married	1364 (62.1%)	612 (27.9%)	
Single	191 (8.7%)	29 (1.3%)	
Drinking			0.0562
No	448 (20.4%)	211 (9.6%)	
Yes	1107 (50.4%)	430 (19.6%)	
Stress			0.0004
Extremely	57 (2.6%)	21 (1%)	
Very	312 (14.2%)	104 (4.7%)	
Slightly	904 (41.2%)	350 (15.9%)	
Rarely	282 (12.8%)	166 (7.6%)	
Smoking			<0.0001
Former and Never	984 (44.8%)	475 (21.6%)	
Current	571 (26%)	166 (7.6%)	

Continuous and categorical features are presented as the mean (standard deviation) and frequency (percentage). AST_Weekdays: average sleep time per day on weekdays, AST_Weekends: average sleep time per day on weekdays

Table 2. Sociodemographic Features and Statistical Analysis in Women

Feature	Normal (n=2117)	Hypertension (n=782)	p-value
Age	51.06 (12.66)	67.04 (10.44)	<0.0001
Systolic BP	114.1 (16.57)	130.6 (16.93)	<0.0001
Diastolic BP	73.94 (9.598)	75.5 (10.95)	0.0005
AST_Weekdays	420.3 (77.17)	425.3 (91.09)	0.1745
AST_Weekends	456.2 (85.21)	438.2 (93.96)	<0.0001
Height	158.2 (6.1)	153.6 (6.214)	<0.0001
Weight	58.05 (9.248)	59.73 (9.683)	<0.0001
Waist circumference	77.92 (8.919)	85.07 (8.88)	<0.0001
Body mass index	23.18 (3.395)	25.27 (3.452)	<0.0001
Income (quintile)			0.0011
Low	499 (17.2%)	214 (7.4%)	
Lower-middle	529 (18.2%)	211 (7.3%)	
Middle	514 (17.7%)	200 (6.9%)	
Upper-middle	575 (19.8%)	157 (5.4%)	
Education			<0.0001
Elementary school or less	320 (11%)	454 (15.7%)	
Middle school	217 (7.5%)	115 (4%)	
High school	709 (24.5%)	158 (5.5%)	
University or higher	871 (30%)	55 (1.9%)	
Occupation			<0.0001
Professionals and expert	333 (11.5%)	15 (0.5%)	
Clerks	228 (7.9%)	15 (0.5%)	
Service and sale workers	363 (12.5%)	117 (4%)	
Agricultural and fishery workers	39 (1.3%)	39 (1.3%)	
Device and assembly workers	72 (2.5%)	17 (0.6%)	
Elementary occupations	198 (6.8%)	105 (3.6%)	
Unemployed	884 (30.5%)	474 (16.4%)	
Household members			<0.0001
1	215 (7.4%)	210 (7.2%)	
2	580 (20%)	315 (10.9%)	
3	552 (19%)	137 (4.7%)	
4	554 (19.1%)	69 (2.4%)	
5	156 (5.4%)	42 (1.4%)	
6 and more	60 (2.1%)	9 (0.3%)	
House type			<0.0001
Detached house	549 (18.9%)	368 (12.7%)	
Apartment	1291 (44.5%)	314 (10.8%)	
Multifamily house	263 (9.1%)	95 (3.3%)	
Etc.	14 (0.5%)	5 (0.2%)	
Marital status			<0.0001
Married	1983 (68.4%)	766 (26.4%)	
Single	134 (4.6%)	16 (0.6%)	
Drinking			<0.0001
No	1189 (41%)	586 (20.2%)	
Yes	928 (32.0%)	196 (6.8%)	
Stress			<0.0001
Extremely	107 (3.7%)	36 (1.2%)	
Very	466 (16.1%)	147 (5.1%)	
Somewhat	1216 (41.9%)	376 (13%)	
Rarely	328 (11.3%)	223 (7.7%)	
Smoking			0.1320
Former and Never	2001 (69%)	750 (25.9%)	
Current	116 (4%)	32 (1.1%)	

Continuous and categorical features are presented as the mean (standard deviation) and frequency (percentage). AST_Weekdays: average sleep time per day on weekdays, AST_Weekends: average sleep time per day on weekdays

제시된 변수들 중에서 systolic BP와 diastolic BP는 고혈압 진단에 직접적으로 활용되기 때문에 통계적 유의성 검증에서만 사용되어졌고, 실제 고혈압 예측 모델링에서는 제외되었다. Age, systolic BP, diastolic BP, height, waist circumference, body mass index, education level, occupation, household members, house type, marital status, smoking 변수들은 정상군/고혈압 샘플들 간에서 통계적 유의성이 매우 높았다 (p = 〈0.0001). 또한 AST_Weekends와 stress 변수도 통계적 유의성이 존재하였다. 그러나, AST_Weekdays, weight, income, drinking은 정상군/고혈압군 사이에서 통계적 유의성이 나타나지 않았다.

여성 Table 2에서는 age, systolic BP, AST_Weekends, height, weight, waist circumference, body mass index, education level, occupation, household members, house type, marital status, drinking, stress 변수들이 정상군/고혈압군 사이에서 통계적 유의성이 매우 높았다(p = 〈 0.0001). 그 외에도 diastolic BP와 income 변수도 통계적 유의성이 높았으나, AST_Weekdays와 smoking은 정상군/고혈압군 샘플들 사이에서 통계적 유의성이 나타나지 않았다. 남녀간 비교에서, 남성에서는 AST_Weekdays, weight, income, drinking 변수만이 통계적 유의성이 나타나지 않았고, 여성에서는 AST_Weekdays와 smoking에서 통계적 유의성이 나타나지 않았다.

3.2 고혈압 예측 모델 성능 평가

Fig. 1과 2는 남녀 각각에 대한 고혈압 예측모델의 주요 성능평가 결과를 나타낸다. 전체적으로 남성보다는 여성에서의 고혈압 예측성능이 높은 것으로 나타났다.

남성에서는 wrapper 메소드를 이용한 모델이 full feature set 및 CFS 메소드를 이용한 모델과 비교하여 가장 높은 예측 성능을 나타내었다 (ROC = 0.790, kappa = 0.396). 그러나 남녀 모두에서 세 가지 모델의 예측성능 차이는 다소 낮은 것으로 나타났다. Wrapper 메소드 기반 예측 모델에서 사용

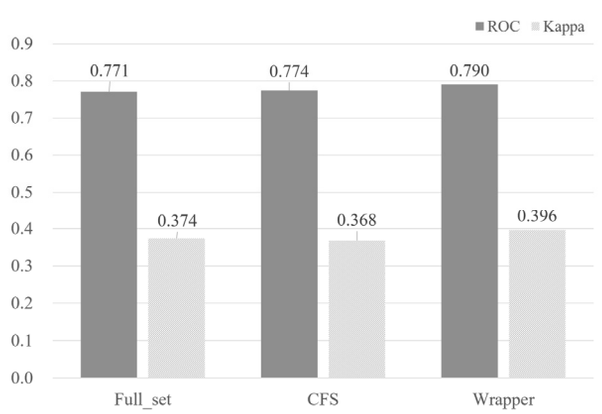

Fig. 1. Prediction Performance Evaluation of Hypertension in Men (ROC: The Area Under the Receiver Operating Characteristic Curve, Full Set: Full Feature Set, CFS: Correlation-Based Feature Subset Method, Wrapper: Wrapper-Based Feature Subset Method)

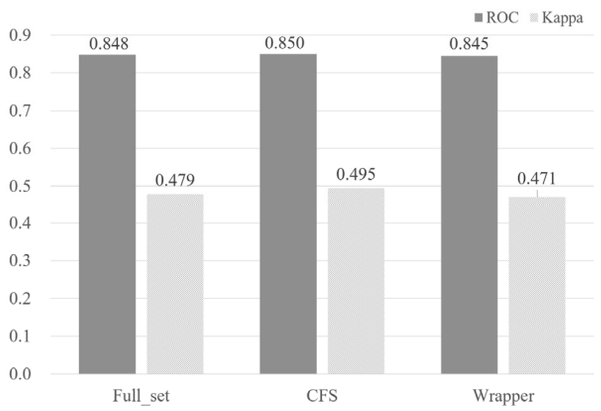

Fig. 2. Prediction Performance Evaluation of Hypertension in Women (ROC: The Area Under the Receiver Operating Characteristic Curve, CFS: Correlation-Based Feature Subset)

된 최종 변수들은 age, AST_Weekdays, waist circumference, body mass index, income, smoking으로 총 6개의 변수들이 포함되었고, 세부적인 성능평가에서는 고혈압군에서 0.487의 sensitivity, 0.117의 1-specificity를 나타내었고, 정상군에서는 0.883의 sensitivity, 0.513의 1-specificity를 나타내었다.

여성에서는 CFS 메소드를 이용한 모델이 가장 높은 예측 성능을 나타내었다 (ROC = 0.850, kappa = 0.495). 이 모델에 사용된 최종 변수들은 age, waist circumference, education으로써 총 3개의 변수가 채택되었다. 세부 성능평가에서는 고혈압군에서 0.696의 sensitivity, 0.175의 1-specificity를 나타내었고, 정상군에서 0.825의 sensitivity, 0.304의 1-specificity를 나타내었다. 생성된 모델들에 대한 세부적인 성능평가 결과는 Table 3에 제시하였고, 최종 모델에 사용된 변수들의 리스트는 Table 4에 제시하였다.

Table 3. Detailed Performance Evaluation of Each Model

Gender	Method	Class	Sen.	1-spe.	F-me.	MCC
Men	Full set	Hyper.	0.618	0.226	0.571	0.376
		Normal	0.774	0.382	0.802	
	CFS	Hyper.	0.61	0.225	0.566	0.37
		Normal	0.775	0.39	0.801	
	Wrapper	Hyper.	0.487	0.117	0.55	0.403
		Normal	0.883	0.513	0.843	
Women	Full set	Hyper.	0.721	0.202	0.636	0.486
		Normal	0.798	0.279	0.84	
	CFS	Hyper.	0.696	0.175	0.642	0.498
		Normal	0.825	0.304	0.852	
	Wrapper	Hyper.	0.565	0.113	0.604	0.473
		Normal	0.887	0.435	0.866	

Sen.: sensitivity, 1-spe.: 1-specificity, F-me.: F-measure, MCC: Mathew correlation coefficient, Hyper.: hypertension, CFS: correlation-based feature subset, Wrapper: wrapper-based feature subset

Table 4. Feature Subset Used in Final Models

	Men			Women		
Feature	Full	CFS	Wra.	Full	CFS	Wra.
Age	○	○	○	○	○	○
AST_Weekdays	○		○	○		○
AST_Weekends	○			○		
Height	○	○		○		
Weight	○			○		
Waist circumference	○	○	○	○	○	
Body mass index	○		○	○		○
Income	○		○	○		
Education	○	○		○	○	
Occupation	○	○		○		
Household members	○	○		○		
House type	○			○		
Marital status	○			○		
Drinking	○			○		
Stress	○			○		
Smoking	○		○	○		
Total number	16	6	6	16	3	3

Wra.: wrapper-based feature selection, AST_Weekdays: average sleep time per day on weekdays, AST_Weekends: average sleep time per day on weekdays, CFS: correlation-based feature subset

4. 토 론

고혈압은 현대인에게 나타나는 비만의 증가, 식습관 및 생활습관의 변화 등에 따라 그 유병율이 매우 증가하는 추세이다[8]. 이러한 중요 질병인 고혈압을 사전에 인지하고 예측하기 위하여 인공지능, 머신러닝, 데이터마이닝 등의 분야에서 수많은 고혈압 예측 또는 식별 모델들이 제시되었다. 고혈압 예측모델에 대한 최신연구들로, 인체계측정보를 이용한 waist circumference, waist-to-height ratio, neck-to-waist ratio 등의 변수들을 기반으로 logistic regression 및 naïve Bayes와 wrapper 메소드를 이용하여 고혈압/저혈압 예측 모델을 제시하였고 남성에서는 0.652의 ROC, 여성에서는 0.721의 ROC 값을 얻었다[12]. 또한 손목의 맥파에서 추출한 변수들을 이용하여 고혈압을 예측하는 모델이 제시되었는데, naïve Bayes와 wrapper 메소드를 이용하여 0.779의 ROC 값을 획득하였고, 이러한 ROC 값은 머신러닝 최신 알고리즘인 least absolute shrinkage and selection operator (LASSO)를 이용한 모델과의 비교에서 보다 우수한 성능을 나타내었다[13]. 또 다른 선행연구에서는 사람 얼굴의 특정 부위들의 색상정보과 나이 및 body mass index를 기반으로 LASSO 알고리즘을 통하여 고혈압 예측 모델을 제시하였는데, 남성에서 0.727 ROC 값과 여성에서는 0.829의 ROC 값을 도출하였다[14]. Logistic regression 알고리즘을 기반으로 한 연구에서는 아홉 개의 최적 변수들의 조합(gender, age, height, weight, triglyceride, LDL 및 HDL cholesterol, uric acid, creatinine)으로 0.758의 ROC 값을 획득하였다[3]. Heo와 Ryu의 선행연구에서는 폐활량정보, 인체계측정보, 혈액정보로부터 추출된 변수들의 조합을 기반으로 logistic regression과 wrapper 메소드를 이용하여 남성에서 0.777 및 여성에서 0.845의 ROC값을 제시하였다[6]. 보다 새로운 임상학적 특징을 활용한 연구에서는 망막안저사진(Retinal fundus photography)에서 추출한 변수들과 사회인구학적 정보를 통합하여 neural network 기반 고혈압 예측모델을 제시하였으며, 이 모델은 고혈압 예측에서 0.766의 ROC 값을 획득하였다[15]. AlKaabi et al.의 연구에서는 사회인구학적 정보뿐만 아니라, 과일 및 채소의 영양분 섭취, 부모의 과거병력 등에 대한 정보를 바탕으로 random forest, decision tree, logistic regression 알고리즘들을 이용하여 고혈압 예측 모델을 제시하였고, 0.799-0.869의 ROC 값을 획득하였다[16].

그러나 이러한 선행연구들은 고혈압 예측 또는 식별모델을 위하여 매우 복잡한 변수 추출방법, 침습적 정보인 혈액정보들, 변수 생성을 위한 특정 고가장비의 필요성 등의 제한점을 지니고 있다. 따라서 이러한 연구들에서 사용된 변수들은 시간 소모적이고 (time-consuming), 고가이며 (expensive), 복잡하고 (complex), 침습적인 (invasive) 형태의 변수들을 바탕으로 개발된 고혈압 모델이므로 실제 국민들의 실생활에서는 적용이 매우 어렵다. 이러한 선행연구들의 모델과 비교하여 본 연구에서 제시한 모델은 매우 단순한 측정변수 또는 기본적인 사회인구학적 정보만을 활용한 고혈압 모델을 제시함에 따라 매우 심플하고, 비용이 거의 없고, 비침습적인 변수들을 기반으로 개발되었다. 또한 기존 연구들에서 제시한 모델들과 본 연구에서 제시한 모델의 예측 성능 비교평가에서도 비록 성별, 나이, 국가, 인종, 지역, 샘플 수, 머신러닝 기법 등에 대한 차별성은 존재하나, 본 연구에서 제시한 모델의 성능이 기존 모델들보다 다소 우수하거나 유사한 성능을 나타내므로 큰 장점이 있다는 것을 알 수 있다.

5. 결 론

전 세계적으로 많은 사람들이 자신의 고혈압을 인지하지 못하거나 치료를 받지 않는 경향이 있다. 본 연구에서는 일반인들이 접근하기 매우 쉬운 사회인구학적 정보만을 바탕으로 고혈압 예측 모델을 제시하였다. 기존 연구들에서 제시한 모델들과는 다르게, 매우 저렴하고, 비침습적이고, 단순한 변수들만을 사용하여 개발되었고, 기존 선행연구에서 제시한 고혈압 모델들의 예측 성능과 비교하여 다소 우수하거나 유사한 성능의 모델을 제시하였다.

본 연구에서 제시한 모델 생성 방법의 활용방안으로는 앱 또는 웹 개발을 통하여 실제 고혈압 환자들뿐만 아니라 일반인도 매우 쉽게 고혈압을 예측 및 관리하는데 사용될 수 있으며, 이러한 모델 개발방법을 확장하여 당뇨, 고지혈증 등의 질병에 대한 예측 모델을 개발할 수 있다. 이러한 모델생성 및 연구결과는 국민 건강에 기여할 수 있을 뿐만 아니라 향후 의학과

전산학의 융합연구의 활성화에 기여할 수 있을 것으로 생각된다. 아울러 머신러닝 및 인공지능 기반 모델들은 고혈압 진단 및 고혈압 관리를 위하여 원격의료 등에서 새로운 전략을 수립할 수 있는 방안을 제시할 수 있을 것으로 판단된다.

References

[1] B. J. Lee and J. Y. Kim, "Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning," *IEEE Journal of Biomedical and Health Informatics*, Vol.20, No.1, pp.39-46, 2015.

[2] B. J. Lee, B. Ku, J. Nam, D. D. Pham, and J. Y. Kim, "Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes," IEEE Journal of Biomedical and Health Informatics, Vol.18, No.2, pp.555-561, 2014.

[3] Z. Ren, et al., "A novel predicted model for hypertension based on a large cross-sectional study," Scientific Reports, Vol.10, No.10615, pp.1-9, 2020.

[4] B. J. Lee and J. Y. Kim, "Identification of the best anthropometric predictors of serum high- and low-density lipoproteins using machine learning," *IEEE Journal of Biomedical and Health Informatics*, Vol.19, No.5, pp.1747-1756, 2015.

[5] B. J. Lee and J. Y. Kim, "Indicators of hypertriglyceridemia from anthropometric measures based on data mining," *Computers in Biology and Medicine*, Vol.57, pp.201-211, 2015.

[6] B. M. Heo and K. H. Ryu, "Prediction of prehypertenison and hypertension based on anthropometry, blood parameters, and spirometry," *International Journal of Environmental Research and Public Health*, Vol.15, No.11, pp.2571, 2018.

[7] B. J. Lee and J. Y. Kim, "Predicting visceral obesity based on facial characteristics," *BMC Complementary and Alternative Medicine*, Vol.14, No.248, pp.1-9, 2014.

[8] K. Tsoi, et al., "Applications of artificial intelligence for hypertension management," *Journal of Clinical Hypertension (Greenwich)*, Vol.23, No.3, pp.568-574, 2021.

[9] B. J. Lee and B. Ku, "A comparison of trunk circumference and width indices for hypertension and type 2 diabetes in a large-scale screening: A retrospective cross-sectional study," *Scientific Reports*, Vol.8, No.13284, pp.1-10, 2018.

[10] B. J. Lee and M. H. Yim, "Comparison of anthropometric and body composition indices in the identification of metabolic risk factors," *Scientific Reports*, Vol.11, No.9931, pp.1-10, 2021.

[11] J. A. Kim, et al., "The prevalence and risk factors associated with isolated untreated systolic hypertension in Korea: The Korean National Health and Nutrition Survey 2001," *Journal of Human Hypertension*, Vol.21, No.2, pp.107-113, 2007.

[12] B. J. Lee and J. Y. Kim, "A comparison of the predictive power of anthropometric indices for hypertension and hypotension risk," *PLoS ONE*, Vo.9, No.1, pp.e84897, 2014.

[13] B. J. Lee, Y. J. Jeon, B. Ku, J. U. Kim, J. H. Bae, and J. Y. Kim, "Association of hypertension with physical factors of wrist pulse waves using a computational approach: A pilot study," *BMC Complementary and Alternative Medicine*, Vol.15, No.222, pp.1-9, 2015.

[14] L. Ang, B. J. Lee, H. Kim, and M. H. Yim, "Prediction of hypertension based on facial complexion," *Diagnostics*, Vol.11, No.540, pp.1-13, 2021.

[15] L. Zhang, et al., "Prediction of hypertension, hyperglycemia and dyslipidemia from retinal fundus photographs via deep learning: A cross-sectional study of chronic diseases in central China," *PLoS ONE*, Vol.15, No.5, pp.e0233166, 2020.

[16] L. A. AlKaabi, L. Sl. Ahmed, M. F. Al Attiyah, and M. E. Abdel-Rahman, "Predicting hypertension using machine learning: Findings from qatar biobank study," *PLoS ONE*, Vol.15, No.10, pp.e0240370, 2020.

이 범 주

https://orcid.org/0000-0003-2682-5716
e-mail : bjlee@kiom.re.kr
2009년 충북대학교 전자계산학과 (공학박사)
2011년 ~ 현 재 한국한의학연구원 디지털임상연구부 책임연구원
관심분야 : Machine Learning & Artificial Intelligence, Data Mining, Public Health, Metabolic Abnormality, Epidemiology