비침습 데이터기반 연령별 대사증후군 위험 예측

강동혁*, 김하민, 정민아**

Non-invasive data-based age-specific metabolic syndrome risk prediction

Kong Donghyuk*, Kim Hamin, Jeong Mina**

요 약

보편적인 헬스케어 장비로 측정할 수 있는 비침습 데이터를 기반으로 대사증후군 발생 위험을 예측하였다. LightGBM 모델을 이용하여 대사증후군 위험 예측을 진행하였고, 입력변수로는 성별, 연령, 허리둘레/신장비율, 체형 지수, 혈압 데이터를 사용하였다. 예측모델의 학습을 위해서 한국 국민건강영양조사(2017~2023년) 데이터를 NCEP ATP III 진단기준에 따라 전체(51872명)을 저위험(38397명), 고위험(13475명) 그룹으로 분류하였다. 연령에 따른 대사증후군 유병률의 연관성을 분석한 결과 연령에 따라 유병률이 상승하는 것을 관찰하였고, 연령층 구분에 따른 대사증후군 위험 예측을 함으로써 연령층별 예측 성능 차이를 분석하였다.

Abstract

We predicted the risk of metabolic syndrome with non-invasive data using healthcare devices. LightGBM was the model used for prediction and sex, age, waist-to-height ratio, body shape index, blood pressure data were the input variables. For training the prediction model, We used Korea National Health and Nutrition Examination Survey(KNHANES 2017~2023) dataset, classifying 51,872 participants into low-risk (38,397) and high-risk (13,475) groups based on NCEP ATP III criteria. An analysis of the relevance between age and the prevalence rate of metabolic syndrome showed that prevalence increases with age. We predicted the risk of metabolic syndrome according to age groups and analyzed the differences in prediction by age groups.

Key words

Non-invasive data, age-specific, WHtR, Metabolic Syndrome, LightGBM, prediction

Ⅰ. 서 론

대사증후군은 심혈관 질환, 당뇨병 등 여러 만성 질환의 위험 요인으로, 현대 사회에서 건강 문제로

*목포대학교 컴퓨공학과, strongstring28@mokpo.ac.kr
목포대학교 컴퓨터공학과, hppol0409@naver.com
**목포대학교 컴퓨터공학과, majung@mnu.ac.kr (교신저자)

※ 본 논문은 2025년 국립목포대학교 글로컬대학 지원에 의하여 연구되었음.

큰 관심을 받고 있다. 대사증후군은 복부 비만, 고혈압, 고혈당, 이상지질혈증 요소가 복합적으로 나타나는 질환으로, 예방과 조기 진단이 중요하다. 대사증후군의 유병률은 비침습 데이터들과 높은 상관관계를 보이고 있으며 이를 바탕으로, 효과적으로 예측하는 모델 개발이 필요하다[1][3]. 기존의 대사증후군 위험 예측 모델들은 주로 의료 기관에서의 검사 데이터를 기반으로 한 방식이었으나, 본 연구에서는 보편적인 헬스케어 장비로 측정할 수 있는 비침습적인 데이터만을 활용하여 대사증후군의 위험을 예측하고자 한다. 이를 위해 의사결정트리 기반 모델을 이용하여, 성별, 연령, 허리둘레, 체중, 신장, 혈압 등 비침습 데이터를 입력 변수로 설정하였다[2][3]. 예측 모델 학습에는 한국 국민건강영양조사(2017~2023) 데이터가 사용되었으며, NCEP ATP III 진단 기준[4]에 따라 데이터를 저위험군과 고위험군으로 분류하였다. 증후군 유병률 차이가 다양하게 나타나는 점을 고려하여, 본 연구는 연령층별로 대사증후군 위험 예측 모델을 학습하고, 이를 전체 예측 모델과 비교 분석하고자 한다. 연령층별 예측 성능 차이를 분석함으로써 각 연령층에 대한 최적화된 대사증후군 예측 모델 개발에 기초자료를 제공할 수 있을 것으로 기대한다[1].

II. 본론

1. 관련 연구

대사증후군 예측과 관련한 머신러닝 기반 연구는 최근 활발히 진행되고 있으며, 대사증후군의 조기 예측과 선행 지표를 밝히기 위해 연령, 성별, 체질량지수, 허리둘레, 혈압, 간 효소 관련 혈액 지표 등 여러 변수를 예측 지표로 사용하여 연구를 진행하였다[3]. 대사증후군 예측에 사용되는 주요 변수 그룹은 체형, 체성분, 심혈관질환 혈액 지표, 간 효소 관련 혈액 지표 및 기타 변수로 나누어지며, 개별변수로써 가장 많이 사용된 변수는 WHtR(허리둘레-신장 비율)이다. 예측에 사용된 분류 모델의 성능 평가는 AUC 값을 주요 지표로 평가함을 확인하였다.

2. 자료원 및 연구 대상

본 연구에서는 한국 국민건강영양조사(2017~2023년, 51872명)데이터 [6]에 표1의 기준에 따라 라벨링을 진행하였다.

기준명	기준
허리둘레(아시아인)	남자 ≥ 90 cm, 여자 ≥ 85 cm
고혈압	축기 혈압 ≥130 mmHg 혹은 이완기 혈압 ≥85 mmHg, 또는 약물치료 중
공복혈당장애	공복 혈당 ≥100 mg/dL 또는 혈당 관리치료 중
고중성지방혈증	triglyceride ≥150 mg/dL 또는 약물치료 중
저 HDL-콜레스테롤	HDL-콜레스테롤 남 <40 mg/dL, 여 < 50 mg/dL 또는 약물치료 중

표 1. NCEP ATP III 진단 기준

3. 입력변수와 학습 모델 설정

본 연구에서는 보편적인 헬스케어 장비로 측정가능한 데이터만으로 대사증후군 위험 예측을 하기 위해서 입력 변수로 SBP(수축기 혈압),DBP(이완기 혈압), ABSI(체형 지수), WHtR(허리둘레-신장 비율), 성별, 연령 6개의 지표를 설정하였다. 입력변수와 대사증후군과의 상관계수는 표2와 같다.

Parameter	correlation coefficient
age	0.383
gender	-0.024
ABSI	0.399
WHtR	0.545
SBP	0.413
DBP	0.296

표 2. MetS에대한 입력변수의 상관계수

ABSI와 WHtR은 체중, 신장, 허리둘레와 같은 신체 계측 지표를 정규화하여 개인의 체형 및 비만 정도를 더욱 정확하게 평가하기 위해 고안된 지표이다. 아래는 ABSI와 WHtR를 계산하는 수식이다.

$$ABSI = \frac{WC}{BMI^{\frac{2}{3}} \times Height^{\frac{1}{2}}} \quad (1)$$

$$WHtR = \frac{WC}{Height} \quad (2)$$

본 연구에서는 학습 이전에 수치형 열에서의 결측치를 처리하기 위한 방법으로 KNN 기법을 사용하여 보정하였다. 보정된 데이터에 대해서 의사결정 트리 기반의 머신러닝 모델 RandomForest, XGBoost, LightGBM을 사용하여 예측하였다. 그림1, 그림2, 그림3을 비교하면 RandomForest, XGBoost 모델은 WHtR의 학습중요도가 0.5 이상이므로 다른 요소보다 편향되어 테스트 시 다른 요소들의 중요도를 무시하는 과적합이 발생한다. 따라서 LightGBM 모델을 예측 모델로 설정하였다.

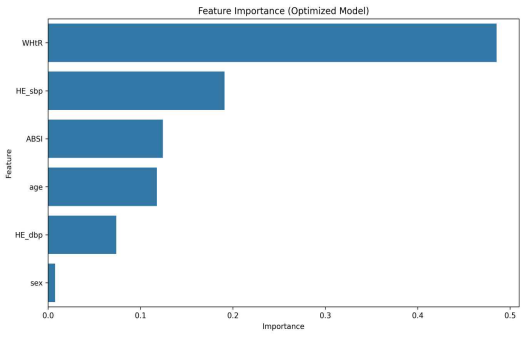

그림 1. RandomForest모델 학습 중요도 그래프

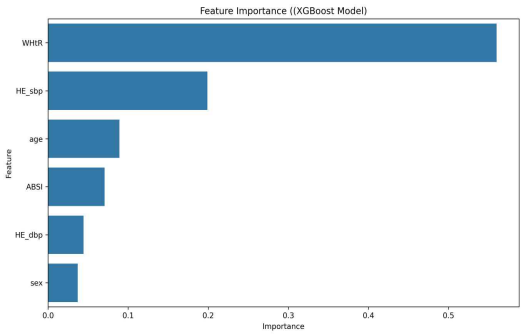

그림 2. XGBoost모델 학습 중요도 그래프

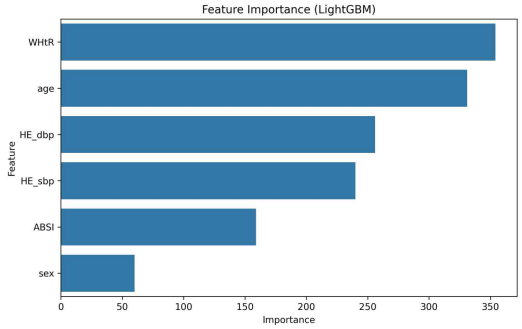

그림 3. LightGBM모델 학습 중요도 그래프

4. LightGBM 모델 구축

LightGBM은 Microsoft에서 개발한 트리 기반의 머신러닝 모델이다. Gradient Boosting Machine의 일종으로 이전 단계에 틀린 부분에 가중치를 더하며 모델을 개선하는 방식으로 학습한다. LightGBM은 Leaf-wise 트리 분할 방식을 사용해서 손실을 최대한 줄이는 리프 노드를 우선적으로 분할하여 예측 오류 손실을 최소화하는 모델이다. 학습에 사용할 LightGBM 모델의 최적 매개변수는 표3이다[7].

Parameters	value
learning_rate	0.1
max_depth	15
n_estimators	100
subsample	0.8

표 3. LightGBM Model Optimized Parameters

5. 연령층 구분 기준 설정 및 분석

본 연구에서는 연령층별 대사증후군 위험 예측을 위해 한국 통계청에서 제공하는 표준화된 구분 기준을 사용하였다[8]. 표4는 학습데이터의 연령층별 대사증후군 유병률을 보여주고 있으며, 표5는 KMeans알고리즘을 이용하여 연령을 제외한 입력 변수들을 군집화시킨 군집별 대사증후군 유병률이다.

Age Range	AgeGroup	population	Prevalence(%)
0-18	Children/Teenager	8945	1.54
19-34	Young Adults	7654	9.96
35-64	Middle-aged Adults	23162	29.24
65+	Older Adults	12111	47.92

표 4. 연령층별 유병률

MetS Cluster Number	MetS False(%)	MetS True(%)
3	98.30	1.70
2	85.26	14.74
1	57.69	42.31
0	37.21	62.79

표 5. 군집별 유병률

표4와 표5를 비교하면 Children/Teenager 그룹과 3번 군집의 유병률이 비슷하며, 그림4에서 제시한 바와

같이 Children/Teenager 그룹의 비율이 가장 큰 것을 확인할 수 있다. 나머지 군집 또한 표3의 연령층 유병률이 상관관계를 가지는 것을 확인할 수 있다.

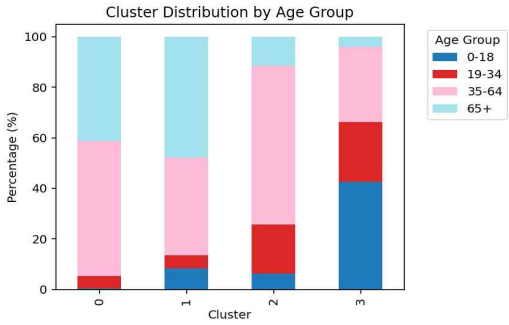

그림 4. 군집별 연령 비중 그래프

연령별로 LightGBM 모델로 학습시킨 모델의 성능은 표6에서 제시하였다. 연령층이 높아질수록 f1-score와 recall(재현율)은 높아지고, AUC는 낮아진다. 따라서 본 연구에서의 대사증후군 위험 예측 모델은 연령층이 높아질수록 실제 예측 성능은 향상되고, 연령층이 낮을수록 실제 예측성능은 하락하지만, 모델의 지표, 최적화를 통해 실제 성능이 향상될 잠재성이 높음을 확인 할 수 있다.

Age range	Accuracy	Recall	F1-score	AUC
Children, Teenager	0.986	0.185	0.286	0.936
Young Adult	0.912	0.362	0.451	0.909
Middle Adult	0.807	0.605	0.647	0.870
Older Adult	0.732	0.767	0.733	0.794

표 8. 연령층별 LightGBM모델 성능표

III. 결 론

본 논문은 보편적인 헬스케어 장비로 측정할 수 있는 비침습 데이터를 기반으로 대사증후군 발생 위험을 예측하였다. 앙상블 학습 중 LightModel을 이용하여 대사증후군 위험 예측을 진행하였고, 연령에 따른 대사증후군 유병률의 연관성을 분석한 결과 연령에 따라 유병률이 상승하는 것을 관찰하였고, 연령층 구분에 따른 대사증후군 위험 예측을 함으로써 연령층별 예측 성능 차이를 분석하였다.

ACKNOWLEDGMENT

본 논문은 2025년 국립목포대학교 글로컬대학 지원에 의하여 연구되었음.

참 고 문 헌

[1] C. Weng, H. Yuan, X. Tang, Z. Huang, K. Yang, W. Chen, P. Yang, Z. Chen and F. Chen, “Age- and gender dependent association between components of metabolic syndrome and subclinical arterial stiffness in a Chinese population,” I
DOI: https://www.medsci.org/v09p0730.htm

[2] B. H. Yu, A. R. Choi and T. H. Kim, “Prediction of metabolic syndrome using the CatBoost model,” Journal of the Korea Academia-Industrial Cooperation Society, vol. 25, no. 4, pp. 324-332, April 2024.
DOI: https://doi.org/10.5762/KAIS.2024.25.4.324

[3] D. Seong, K. Jeong, S. Lee and Y. Baek, “Metabolic syndrome prediction model for Koreans in recent 20 years: a systematic review,” The Journal of the Korea Contents Association, vol. 21, no. 8, pp. 662-674, August 2021.
DOI: https://doi.org/10.5392/JKCA.2021.21.08.662

[4] S. N. Grundy, D. Becker, L. T. Clark, R. S. Cooper, M. A. Denke, W. J. Howard et al, “Third report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) Final Report“ Circulation, pp.3143-3421, Vol.106, No.25, Dec. 2002.

[5] J. H. Kim, K. H. Kim, G. B. Park and W. K. Lee, “Metabolic syndrome prevalence change before and after the COVID-19 epidemic: Using data from the National Health and Nutrition Examination Survey 2018-2021,” Journal of Health Informatics and Statistics, vol. 49, no. 4, pp. 315-

324, 2024.
DOI: https://doi.org/10.21032/jhis.2024.49.4.315

[6] Korea Disease Control and Prevention Agency (KDCA), Korea National Health and Nutrition Examination Survey (KNHANES) Raw Data Download.
[Online]. Available: https://knhanes.kdca.go.kr/knhanes/rawDataDwnld/rawDataDwnld.do, [Accessed: May 15, 2025].

[7]G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," Advances in Neural Information Processing Systems, vol. 30, 2017.

[8] Statistics Korea, Statistics Me Service, SGIS.
[Online]. Available: https://sgis.kostat.go.kr/view/statsMe/statsMeMain#1, [Accessed: May 15, 2025].