Journal of KIIT. Vol. 24, No. 4, pp. 49-61, Apr. 30, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.4.49

당뇨 예측모델 개발에서 성별 및 알고리즘에 따른 변수 선택 차이와 예측 성능 비교

김정희*[1], 김유빈*[2], 서주희*[3], 정재효**[1], 박주용**[2]

Comparison of Sex- and Algorithm-Dependent Differences in Variable Selection and Predictive Performance for Diabetes Prediction Model Development

Jeong-Hee Kim*[1], Yu-Bin Kim*[2], Ju-Hee Seo*[3], Jaehyo Jung**[1], and JooYong Park**[2]

이 논문은 2025학년도 을지대학교 학술연구비 지원에 의하여 이루어진 것임(EJRG-25-21)

요 약

본 연구는 성별 및 알고리즘에 따른 미진단 당뇨 예측모델의 변수 선택 차이와 영향을 비교하였다. 남성 36,558명과 여성 70,308명의 데이터에서 19개의 비침습적 변수를 사용하였고 통계적인 방법과 머신러닝 모델(Random Forest, Logistic Regression, AdaBoost, XGBoost, LightGBM)을 적용하여 구축한 미진단 당뇨 예측모델을 비교하였다. 분석 결과 남성은 맥박수와 체지방률, 여성은 체성분 지표가 주요 변수로 도출되었으며, 동일 성별 내에서도 알고리즘에 따라 최종 변수 구성과 순위가 상이함을 확인하였다. 예측 성능(AUC)은 여성(0.77 – 0.80)이 남성(0.70 – 0.72)보다 높아 성별 특성의 영향력을 입증하였다. 본 연구는 성별 분리 모델링과 알고리즘 선택이 당뇨 예측모델의 해석력과 성능 향상에 중요한 고려 요소임을 제시한다.

Abstract

This study compared sex- and algorithm-dependent differences in variable selection and their impact on performance in diabetes prediction modeling. Undiagnosed diabetes was defined in participants without a prior physician diagnosis as fasting plasma glucose ≥126 mg/dL or glycated hemoglobin (HbA1c) ≥6.5%. The final sample included 36,558 men and 70,308 women, and 19 predictors covering demographics, disease/family history, lifestyle, and non-invasive clinical indicators were analyzed. Sex-stratified models were built using statistical logistic regression with stepwise selection and machine learning models (Random Forest, Logistic Regression, AdaBoost, XGBoost, and LightGBM). In men, pulse rate and percent body fat were ranked highly, whereas in women, body composition indicators were more prominent. Variable composition and ranking also differed across algorithms within each sex. Predictive performance was generally higher in women than in men (AUC: 0.77 – 0.80 vs. 0.70 – 0.72). These findings indicate that sex stratification and algorithm choice influence variable composition and performance interpretation.

Keywords

undiagnosed diabetes, sex-specific model, machine learning, statistical model, variable selection, prediction

* 을지대학교 빅데이터의료융합학과
- ORCID[1]: https://orcid.org/0009-0006-4157-263X
- ORCID[2]: https://orcid.org/0009-0009-0948-5685
- ORCID[3]: https://orcid.org/0009-0009-5515-1395
** 을지대학교 첨단학부 빅데이터인공지능전공(**[2] 교신저자)
- ORCID[1]: https://orcid.org/0000-0003-1852-3267
- ORCID[2]: https://orcid.org/0000-0002-6444-3754

· Received: Mar. 10, 2026, Revised: Apr. 16, 2026, Accepted: Apr. 19, 2026

· Corresponding Author: JooYong Park
Dept. of Big Data and Artificial Intelligence, Eulji University
Korea
Tel.: +82-31-740-7160, Email: judepark0501@gmail.com

Ⅰ. 서 론

질병 예측모델은 오랫동안 로지스틱 회귀분석과 같은 통계 기반 모델을 중심으로 발전해 왔다. 통계 기반 접근은 결과와 예측변수 간의 관계를 회귀계수로 추정하고, 그 크기와 방향을 통해 해석 가능한 결과를 제공한다[1]. 그러나 보건의료 데이터의 규모가 커지고 고차원으로 확장되면서, 변수 간 상호작용과 비선형 구조가 빈번해졌고 단순한 선형 관계를 전제로 한 전통적 접근만으로는 복잡한 패턴을 충분히 반영하기 어려워졌다[2][3]. 이러한 변화는 데이터 기반으로 패턴을 학습해 일반화 성능을 높일 수 있는 머신러닝 기법의 활용 증가로 이어지고 있다[2][3].

통계적 예측모델과 머신러닝 예측모델은 모두 예측을 목적에 두고 있지만, 모델 구축의 전제, 학습 방식, 해석 가능성 측면에서 관점의 차이가 존재한다. 통계적 방법은 해석 가능성과 안정성을 강점으로 가지는 반면, 머신러닝은 다양한 알고리즘을 통해 복잡한 비선형 구조를 유연하게 반영하는 데 유리하다[2][4]. 다만 통계적인 방식이 주로 관계의 설명과 해석을 중시하는 반면, 머신러닝 기법은 상대적으로 예측 성능의 최적화에 초점이 맞춰질 수 있어, 연구 목적에 따라 비교 기준이 달라질 수 있다[4][5]. 또한 임상 분야에서 예측모델 비교 문헌을 종합한 연구에서는 머신러닝이 로지스틱 회귀분석보다 일관되게 우수하다고 보기 어렵고, 비교 설계와 검증 과정의 차이가 성능 추정에 영향을 줄 수 있음을 언급하였다[6]. 따라서 예측모델 비교는 성능지표의 차이뿐 아니라, 어떤 정보가 어떤 방식으로 반영되어 최종 모델이 구성되는지까지 함께 검토할 필요가 있다.

당뇨병은 2022년 국민건강영양조사에서 19세 이상 성인의 유병률이 남성 11.2%, 여성 6.9%로 보고되었다[7]. 또한 대한당뇨병학회 2022년 보고서도 남녀 유병률의 차이와 연령대별 양상이 다름을 제시하였으며, 이는 당뇨 예측에 있어서 성별 맞춤형 모델이 필요함을 뒷받침한다[8]. 당뇨병은 공복혈당과 당화혈색소와 같은 혈액검사로 진단되지만, 공복상태 유지와 채혈은 검사 과정의 부담으로 작용할 수 있다. 이로 인해 건강검진 및 설문 기반의 비침습적 변수만으로 미진단 당뇨를 선별하고자 한 연구들이 보고되어 왔다. 국민건강영양조사 자료를 활용해 비침습적 변수로 미진단 당뇨를 예측하고 통계모델과 머신러닝 모델의 성능을 비교한 연구가 제시되었으나, 성별을 층화하여 변수 선택 구조의 차이를 비교하거나 성별에 따른 모델 구성 차이를 중심 질문으로 다루지는 않았다[9]. 또한 비침습적 변수만을 활용해 중장년 여성에서 당뇨 및 공복혈당장애 분류를 시도한 연구도 있으나, 여성 집단에 한정된 예측모델이라는 성격을 지니며 표본 규모가 제한되고 남성에 대한 분석이 포함되지 않았다는 한계가 남는다[10]. 이처럼 비침습적 변수를 이용한 당뇨 예측 연구는 축적되고 있지만, 성별을 분리한 환경에서 통계적 방법과 머신러닝 알고리즘별 변수 선택 결과가 어떻게 달라지고 그 차이가 성능 차이로 어떻게 연결되는지에 대한 비교는 여전히 충분히 제시되지 않았다.

이에 본 연구는 성별 분리 환경에서 통계 기반 방법과 머신러닝 기법을 적용하여 변수 선택 과정과 최종 변수 구성의 차이를 비교하고, 그 차이가 예측 성능 및 해석 가능성에 미치는 영향을 체계적으로 확인하고자 한다. 동일한 자료와 동일한 절차를 적용하더라도 성별에 따라 선택되는 변수의 종류와 순위가 달라질 수 있는지, 또한 알고리즘 선택이 변수 구성과 예측 성능에 어떤 영향을 미치는지에 대해 체계적으로 제시하고자 한다.

본 논문의 구성은 다음과 같다. 제2장 연구방법에서는 KoGES HEXA 코호트 자료를 활용한 연구 대상자 선정 과정과 미진단 당뇨 및 19개 예측변수의 정의를 기술하고, 통계적 방법(Stepwise selection)과 머신러닝 기법 RFE(Recursive Feature Elimination) FI(Feature Importance)을 활용한 성별 분리 모델링 및 평가 절차를 설명한다. 제3장 결과에서는 성별 및 당뇨 여부에 따른 대상자 특성 비교와 함께, 알고리즘별 변수 선택 결과 및 예측 성능의 차이를 제시한다. 마지막으로 제4장 고찰 및 결론에서는 분석 결과를 바탕으로 성별 특화 예측모델링의 학술적, 보건학적 의의를 논의하고 최종 결론을 도출한다.

II. 연구방법

2.1 연구대상

본 연구에서는 한국인유전체역학조사사업(KoGES, Korean Genome and Epidemiology Study)의 도시 기반 코호트(HEXA, Health Examinees Study) 자료를 사용하였다[11]. KoGES는 한국인에서 유전 및 환경 요인이 주요 만성질환의 발생에 미치는 영향을 장기 추적하여 규명하기 위해 정부 주도로 구축된 대규모 전향적 코호트 컨소시엄이다. 그 중 HEXA는 국가건강검진 수검자 등록부를 기반으로 40세 이상 성인 남녀를 모집하며, 숙련된 조사원이 표준화된 설문과 신체검진을 수행하고 2~4년 간격으로 추적 조사를 진행하였다[11]. HEXA 기반조사에 참여한 총 173,357명 중, 기반조사 시점에서 당뇨병 과거력이 있는 11,503명을 제외하였다. 또한 인구학적 특성, 생활습관 변수 및 비침습적 임상지표에서 결측치가 확인된 54,282명을 추가로 제외하였다. 그 결과 최종 분석 대상자는 총 106,866명이었다. 이 중 남성은 36,558명으로 4.4%가 당뇨병으로 분류되었으며, 여성은 70,308명으로 2.4%가 당뇨병으로 분류되었다.

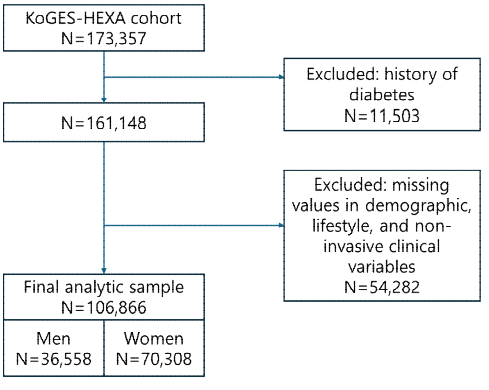


그림 1. 연구대상자 흐름도
Fig. 1. Flow chart of the study population

본 연구는 을지대학교 기관생명윤리위원회(Institutional Review Board, EUIRB2024-105)의 심의를 거쳐 승인되었으며, 관련 연구 윤리 지침을 준수하여 수행되었다.

2.2 결과변수

본 연구에서는 미진단 당뇨(Undiagnosed diabetes)를 결과변수로 하며, 기반조사 시점의 공복혈당과 당화혈색소를 이용하여 정의하였다. 기반조사 시점에서 과거에 의사로부터 당뇨 진단을 받은 적이 없는 대상자들 중 공복혈당이 126mg/dl 이상이거나 당화혈색소가 6.5% 이상인 경우 미진단 당뇨로 분류하였다. 이는 의사 진단을 받은 이후의 생활습관 변화 또는 비침습적 임상지표 변화가 예측변수에 반영되어 발생할 수 있는 역인과 가능성을 최소화하기 위함이다.

2.3 예측변수

예측변수는 기반조사 시점에서 활용 가능한 인구학적 특성, 질병력 및 가족력, 생활습관, 비침습적 임상지표를 포함한 총 19개 변수로 구성하였다.

2.3.1 인구학적 변수 및 병력, 가족력

나이는 만 연령으로 수집된 값을 사용하였다. 교육 수준은 최종학력을 기준으로 응답된 자료를 중학교 졸업이하, 고등학교 졸업, 대학교 졸업 이상으로 재분류하여 사용하였다. 현재 결혼 상태는 미혼, 기혼, 별거, 이혼, 사별 동거, 기타로 조사되었고, 기혼과 그 외 그룹으로 분류하여 사용하였다. 가정의 월평균 수입은 “귀하 가정의 월 평균 수입(가구전체소득)은 어느 정도 되십니까?” 라는 문항으로 수집되었고, 200만원 미만, 200~400만원, 400만원 이상으로 재분류하여 사용하였다. 고혈압 과거력 여부와 당뇨병 가족력 유무는 예/아니오 응답 그대로 활용하였다.

2.3.2 생활습관 변수

흡연은 평생 비흡연, 과거 흡연, 현재 흡연 3개 범주로 조사되었으며, 분석에서는 평생 흡연 노출 여부를 기준으로 평생 비흡연과 흡연 경험자(과거 흡연 및 현재흡연)로 재분류하여 사용하였다. 음주

역시 유사한 방식으로 조사되었으며, 평생 비음주자와 음주 경험자(과거음주 및 현재음주)로 재분류하여 예측변수로 활용하였다.

신체활동의 경우, 평소에 땀이 날 정도의 운동 여부를 먼저 조사한 뒤 운동을 한다고 응답한 대상자에 대해 주당 평균 운동 빈도와 1회 평균 운동시간을 추가로 조사하였다. 이를 이용하여 식 (1)과 같이 주당 평균 운동시간을 산출하였다.

$$\text{주당평균운동시간}(Minutes/Week) = \text{주당평균운동횟수} \times \text{1회평균운동시간}(Minutes) \quad (1)$$

걷기의 경우 지난 1년 동안 산보 또는 산책 활동 여부를 조사한 뒤, 활동을 하는 대상자에 대해 주당 평균 걷기 횟수와 1회 평균 걷기시간을 추가로 조사하였다. 식 (2)와 같이 운동과 동일한 방식으로 주당 평균 걷기시간을 산출하였다.

$$\text{주당평균걷기시간}(Minutes/Week) = \text{주당평균걷기횟수} \times \text{1회평균걷기시간}(Minutes) \quad (2)$$

산출된 주당 운동시간과 주당 걷기시간은 WHO 권고 기준을 참고하여 0분(실천 안함), 150분 미만, 150분 이상으로 범주화하였다.

2.3.3 혈압 및 신체계측

비침습적 임상지표만을 활용하기 위하여 본 연구에서는 수축기혈압, 이완기혈압, 맥박수, 체질량지수, 체지방률, 체지방량, 근육량, 내장지방량, 복부비만율(허리-엉덩이둘레비)을 사용하였다. 맥박수는 표준 절차에 따라 30초 또는 1분 동안 측정하였고, 수축기혈압과 이완기혈압은 표준화된 수은혈압계를 이용하여 2회 측정한 뒤 평균값을 분석에 사용하였다. 허리둘레와 엉덩이둘레는 직접 계측값을 사용하였으며, 복부비만율은 허리둘레를 엉덩이둘레로 나눈 값으로 산출하였다. 체질량지수는 신장과 체중의 실측값으로 계산하였고, 체지방률, 체지방량, 근육량 및 내장지방량 등의 체성분 지표는 다주파수 생체전기저항분석법(MF-BIA, Multifrequency Bioelectrical Impedance Analysis) 기반의 체성분분석기(InBody 3.0; Biospace, Seoul, Korea)로 측정된 값을 사용하였다[11].

2.4 머신러닝 및 통계 방법 예측모델링

본 연구는 성별에 따른 변수 선택 결과와 예측 성능의 차이를 비교하기 위해 남성과 여성 자료를 분리한 후 동일한 모델링 절차를 적용하였다. 모델링 과정은 데이터 분할, 변수 선택, 모델 학습 및 교차검증, 최적 임계값 설정, 최종 성능 평가의 순서로 수행하였다.

각 성별 자료는 미진단 당뇨의 비율을 유지하기 위해 학습, 검증, 평가 데이터를 6:2:2로 분할하였으며, 모든 모델에서 동일한 분할과 동일한 입력변수 후보군을 사용하여 비교의 공정성을 확보하였다. 변수 선택은 학습 데이터에서만 수행하여 정보 누수를 방지하였고, 재현성을 위해 시드를 고정하였다. 분석은 R(version 4.5.1)과 Python(version 3.10)을 사용하여 수행하였다.

2.4.1 변수 선택 방법

변수 선택은 통계적 방법과 머신러닝 기법의 특성을 반영하여 서로 다른 절차를 적용하였다. 통계기반 모델에서는 stepwise selection을 적용하여 변수의 추가 및 제거를 반복하면서 최종 변수를 선택하였다. Stepwise selection은 AIC(Akaike Information Criterion)을 기준으로 수행하였다. 머신러닝 모델에서는 RFE과 FI 기반 선택을 적용하였다.

RFE는 전체 변수에서 시작하여 모델 성능에 대한 기여도가 낮은 변수를 반복적으로 제거하는 방식으로 수행하였다. 본 연구에서는 최종 변수 수를 10개와 15개로 설정하여 비교한 후, 교차검증 성능이 더 우수한 10개 변수 구성을 최종 변수 집합으로 선정하였다. FI 기반 선택은 학습된 모델에서 산출된 변수 중요도 순위를 이용하여 상위 변수를 선택하는 방식으로 수행하였으며, 동일한 기준을 적용하여 상위 10개 변수를 최종 변수 집합으로 선정하였다.

2.4.2 예측모델 구축

본 연구에서는 로지스틱 회귀분석을 통계 기반 방법과 머신러닝 기반 비교모델의 두 맥락에서 모두 활용하였다. 먼저 통계 기반 방법에서는 stepwise selection과 결합하여 최종 변수를 선택하고, 추정된 회귀계수를 통해 변수의 방향성과 크기를 해석하는 기준모델로 사용하였다. 반면 머신러닝 기반 비교에서는 로지스틱 회귀분석을 하나의 분류 알고리즘으로 간주하여, 다른 머신러닝 모델과 동일한 학습 및 평가 절차 하에서 예측 성능과 변수 선택 결과를 비교하였다. 즉 동일한 로지스틱 회귀분석이라도 본 연구에서는 해석 중심의 통계적 모델과 예측 중심의 비교 알고리즘이라는 서로 다른 목적에 따라 구분하여 사용하였다.

머신러닝 모델로는 Random Forest, AdaBoost, XGBoost, LightGBM과 함께 로지스틱 회귀분석 기반 분류모델을 포함하였다. 이들 모델은 비선형 관계와 변수 간 상호작용을 반영할 수 있으며, 특히 앙상블 기반 알고리즘을 통해 복잡한 데이터 구조를 유연하게 학습할 수 있다는 특징이 있다. 또한, 학습 과정에서 변수 중요도를 산출할 수 있어 알고리즘별 변수 선택 결과의 차이를 비교하는 데 적합하다. 각 모델은 성별별 자료에 대해 동일한 절차로 학습 및 평가하였으며, 주요 하이퍼파라미터는 교차검증 기반 방법을 통해 설정하였다. 각 모델의 최종 하이퍼파라미터 설정값은 표 1에 제시하였다.

2.4.3 교차검증 및 성능 평가

각 모델은 학습 자료에서 교차검증을 통해 학습 및 튜닝하였으며, 최종 성능 평가는 독립된 평가 자료를 이용하여 수행하였다. 예측 성능 평가는 ROC 곡선 기반 지표를 중심으로 비교하였고, 분류 임계값은 각 모델의 ROC 곡선에서 Youden index를 최대화하는 지점으로 설정하였다. 해당 임계값을 기준으로 민감도, 특이도 및 기타 분류 성능 지표를 산출하였다. 본 연구에서는 단순한 성능 우열 비교뿐 아니라, 성별 및 알고리즘에 따라 선택된 변수 구성이 어떻게 달라지는지를 함께 해석하였다.

III. 결 과

3.1 대상자 특성에 따른 남녀 차이 분포

표 2는 연구 대상자의 성별에 따른 인구사회학적 특성, 생활습관, 및 임상 지표를 비교한 결과를 보여주고 있다. 남성과 여성에서 모든 변수들이 통계적으로 유의미한 차이를 보이고 있으며(p-value <0.0001), 성별에 따른 예측모델 개발의 필요성을 시사한다. 남성은 여성보다 만 나이, 소득, 교육 수준이 높고, 배우자와 함께 사는 비율과 고혈압 유병율도 여성보다 높았다. 당뇨 가족력은 남성보다 여성이 많았고, 평생 비흡연, 평생 비음주자도 여성이 많은 것으로 나타났다. 엉덩이둘레, 혈압, 근육량, 복부지방량, 복부비만율, 체질량지수는 남성이 높고, 체지방률과 맥박은 여성이 높았다. 본 연구의 결과변수인 미진단 당뇨는 남성이 4.4%, 여성이 2.4%로 확인되었다.

표 1. 성별에 따른 머신러닝 모델별 최적 하이퍼파라미터 설정
Table 1. Optimized hyperparameters for machine learning models stratified by sex

Model	Men	Women
XGboost	- n_estimators = 300 - learning_rate = 0.1 - max_depth = 3	- n_estimators = 500 - learning_rate = 0.01 - max_depth = 3
Random Forest	- n_estimators = 1800 - max_depth = 24	- n_estimators = 1800 - max_depth = None
LGBM	- learning_rate = 0.05 - min_child_samples = 40 - reg_lambda = 0.1 - max_depth = 3	- learning_rate = 0.04 - min_child_samples = 20 - max_depth = 3 - reg_lambda = 0.1

표 2. 성별에 따른 특성 비교
Table 2. Comparison of characteristics by sex

	Men (N=36558, 34%)	Women (N=70308, 66%)	p-value*
Age, mean (SD)	53.2 (8.7)	52.1 (7.9)	<0.0001
Income			
<2 million won	9393 (25.7%)	23021 (32.7%)	<0.0001
2-4 million won	16263 (44.5%)	29510 (42.0%)	
≥4 million won	10902 (29.8%)	17777 (25.3%)	
Education			
≤ Middle school	7694 (21.0%)	25499 (36.3%)	<0.0001
High school	12329 (33.7%)	27275 (38.8%)	
≥ College	16535 (45.2%)	17534 (24.9%)	
Marital status			
Living with spouse	34115 (93.3%)	60586 (86.2%)	<0.0001
Living alone	2443 (6.7%)	9722 (13.8%)	
Hypertension			
No	29081 (79.5%)	59148 (84.1%)	<0.0001
Yes	7477 (20.5%)	11160 (15.9%)	
Family history of diabetes			
No	31107 (85.1%)	57036 (81.1%)	<0.0001
Yes	5451 (14.9%)	13272 (18.9%)	
Ever smoke			
No	9893 (27.1%)	67746 (96.4%)	<0.0001
Yes	26665 (72.9%)	2562 (3.6%)	
Ever drink			
No	7038 (19.3%)	45985 (65.4%)	<0.0001
Yes	29520 (80.7%)	24323 (34.6%)	
Exercise			
No	16233 (44.4%)	35684 (50.8%)	<0.0001
< 150 minutes/week	4980 (13.6%)	8558 (12.2%)	
≥ 150 minutes/week	15345 (42.2%)	26066 (37.1%)	
Walk			
No	23889 (65.3%)	42816 (60.9%)	<0.0001
< 150 minutes/week	6340 (17.3%)	12559 (17.9%)	
≥ 150 minutes/week	6329 (17.3%)	14933 (21.2%)	
Hip circumference, mean (SD)	96.1(5.6)	93.6(5.7)	<0.0001
Pulse, mean (SD)	68.5 (9.7)	69.3 (9.2)	<0.0001
Systolic blood pressure, mean (SD)	125.8 (14.4)	120.5 (15.2)	<0.0001
Diastolic blood pressure, mean (SD)	78.8 (9.8)	74.6 (9.8)	<0.0001
Percent of body fat, mean (SD)	23.4 (4.5)	30.6 (4.4)	<0.0001
Muscle mass, mean (SD)	49.0 (5.1)	36.6 (3.5)	<0.0001
Visceral fat mass, median (IQR)	2.4 (1.9 - 3.0)	2.0 (1.5 - 2.5)	<0.0001
Waist-hip ratio, median (IQR)	0.89 (0.86 - 0.92)	0.84 (0.79-0.88)	<0.0001
Body mass index, mean (SD)	24.4 (2.8)	23.6 (2.9)	<0.0001
Undiagnosed diabetes			
No	34951 (95.6%)	68631(97.6%)	< 0.0001
Yes	1607 (4.4%)	1677 (2.4%)	

* Student's t-test was used for normally distributed continuous variables, the Wilcoxon rank-sum test for non-normally distributed continuous variables, and the chi-square test for categorical variables.

3.2 당뇨 여부에 따른 변수 차이 비교

남녀 각각에서 미진단 당뇨 여부에 따라 예측변수의 차이를 확인해 본 결과는 표 3과 표 4와 같다. 남성에서는 결혼상태, 여성에서는 평생 흡연 여부 변수를 제외한 모든 예측변수들에서 통계적으로 유의미한 차이를 보였다 (p-value <0.0001). 남녀 공통적으로 정상 그룹보다 미진단 당뇨인 그룹에서 나이가 많고, 소득이 적었으며, 교육 수준이 낮았다. 고혈압 유병율과 당뇨 가족력도 많았고, 혈압과 모든 비침습적 임상지표들이 유의미하게 높은 현상이 나타났다. 여성에서만 미진단 당뇨 그룹에서 혼자 사는 그룹이 많았고, 평생 비음주 그룹이 많았다. 반면, 남성에서는 미진단 당뇨 그룹에서 음주 경험자와 흡연 경험자가 많았다.

표 3. 남성 대상자의 미진단 당뇨병 유무에 따른 특성 비교
Table 3. Comparison of characteristics by undiagnosed diabetes status in men

	Undiagnosed diabetes		
	No (N=34951, 95.6%)	Yes (N=1607, 4.4%)	p-value*
Age, mean (SD)	53.1(8.7)	54.8 (8.2)	<0.0001
Income			
<2 million won	8962 (25.6%)	431 (26.8%)	0.0028
2-4 million won	15505 (44.4%)	758 (47.2%)	
≥4 million won	10484 (30.0%)	401 (25.0%)	
Education			
≤ Middle school	7293 (20.9%)	401 (25.0%)	< 0.0001
High school	11782 (33.7%)	547 (34.0%)	
≥ College	15876 (45.4%)	659 (41.0%)	
Marital status			
Living with spouse	32622 (93.3%)	1493 (92.9%)	0.5324
Living alone	2329 (6.7%)	114 (7.1%)	
Hypertension			
No	27960 (80.0%)	1121 (69.8%)	<0.0001
Yes	6991 (20.0%)	486 (30.2%)	
Family history of diabetes			
No	29886 (85.5%)	1221 (76.0%)	<0.0001
Yes	5065 (14.5%)	386 (24.0%)	
Ever smoke			
No	9545 (27.3%)	348 (21.8%)	<0.0001
Yes	25406 (72.7%)	1259 (78.3%)	
Ever drink			
No	6760 (19.3%)	278 (17.3%)	0.0458
Yes	28191 (80.7%)	1329 (82.7%)	
Exercise			
No	15449 (44.2%)	784 (48.8%)	0.0004
< 150 minutes/week	4798 (13.7%)	182 (11.3%)	
≥ 150 minutes/week	14704 (42.1%)	641 (39.9%)	
Walk			
No	22781 (65.2%)	1108 (68.9%)	0.0004
< 150 minutes/week	6119 (17.5%)	221 (13.8%)	
≥ 150 minutes/week	6051 (17.3%)	278 (17.3%)	
Hip circumference, mean (SD)	96.0 (5.6)	97.2 (6.2)	<0.0001
Pulse, mean (SD)	68.4 (9.7)	71.7 (10.6)	<0.0001
Systolic blood pressure, mean (SD)	125.4 (14.3)	131.0 (15.0)	<0.0001
Diastolic blood pressure, mean (SD)	78.7 (9.8)	81.2 (9.7)	<0.0001
Percent of body fat, mean (SD)	23.3 (4.4)	25.3 (4.2)	<0.0001
Muscle mass, mean (SD)	49.0 (5.1)	49.6 (5.6)	0.0001
Visceral fat mass, median (IQR)	2.4 (1.9-3.0)	2.8 (2.2-3.5)	<0.0001
Waist-hip ratio, median (IQR)	0.89 (0.86-0.92)	0.91 (0.88-0.94)	<0.0001
Body mass index, mean (SD)	24.3 (2.7)	25.5 (2.9)	<0.0001

* Student's t-test was used for normally distributed continuous variables, the Wilcoxon rank-sum test for non-normally distributed continuous variables, and the chi-square test for categorical variables.

표 4. 여성 대상자의 미진단 당뇨병 유무에 따른 특성 비교
Table 4. Comparison of characteristics by undiagnosed diabetes status in women

	Undiagnosed diabetes		
	No (N=68631, 97.6%)	Yes (N=1677, 2.4%)	p-value*
Age, mean (SD)	52.0 (7.9)	55.9 (7.7)	<0.0001
Income			
<2 million won	22290 (32.5%)	731 (43.6%)	<0.0001
2-4 million won	28862 (42.1%)	648 (38.6%)	
≥4 million won	17479 (25.5%)	298 (17.8%)	
Education			
≤ Middle school	24637 (35.9%)	862 (51.4%)	<0.0001
High school	26698 (38.9%)	577 (34.4%)	
≥ College	17296 (25.2%)	238 (14.2%)	
Marital status			
Living with spouse	59201 (86.3%)	1385 (82.6%)	<0.0001
Living alone	9430 (13.7%)	292 (17.4%)	
Hypertension			
No	58014 (84.5%)	1134 (67.6%)	<0.0001
Yes	10617 (15.5%)	543 (32.4%)	
Family history of diabetes			
No	55861 (81.4%)	1175 (70.1%)	<0.0001
Yes	12770 (18.6%)	502 (29.9%)	
Ever smoke			
No	66144 (96.4%)	1602 (95.5%)	0.0924
Yes	2487 (3.6%)	75 (4.5%)	
Ever drink			
No	44797 (65.3%)	1188 (70.8%)	<0.0001
Yes	23834 (34.7%)	489 (29.2%)	
Exercise			
No	34783 (50.7%)	901 (53.7%)	
< 150 minutes/week	8371 (12.2%)	187 (11.2%)	
≥ 150 minutes/week	25477 (37.1%)	589 (35.1%)	
Walk			
No	41729 (60.8%)	1087 (64.8%)	<0.0001
< 150 minutes/week	12331 (18.0%)	228 (13.6%)	
≥ 150 minutes/week	14571 (21.2%)	362 (21.6%)	
Hip circumference, mean (SD)	93.7 (5.6)	95.7 (6.7)	<0.0001
Pulse, mean (SD)	69.2 (9.2)	72.2 (9.9)	<0.0001
Systolic blood pressure, mean (SD)	120.3 (15.2)	128.6 (16.3)	<0.0001
Diastolic blood pressure, mean (SD)	74.5 (9.7)	78.5 (9.8)	<0.0001
Percent of body fat, mean (SD)	30.5 (4.4)	33.6 (4.1)	<0.0001
Muscle mass, mean (SD)	36.5 (3.5)	37.6 (4.1)	<0.0001
Visceral fat mass, median (IQR)	2.0 (1.5-2.5)	2.6 (2.0-3.2)	<0.0001
Waist-hip ratio, median (IQR)	0.83 (0.79-0.88)	0.88 (0.84-0.92)	<0.0001
Body mass index, mean (SD)	23.5 (2.9)	25.7 (3.4)	<0.0001

* Student's t-test was used for normally distributed continuous variables, the Wilcoxon rank-sum test for non-normally distributed continuous variables, and the chi-square test for categorical variables.

3.3 통계 기반 예측모델과 기계학습 기반 예측모델 비교

3.3.1 변수 선택 및 변수 순위의 차이

표 5와 표 6은 남성과 여성에서 변수 선택 방법 및 모델별 변수 순위를 비교한 결과를 제시한다. 남성과 여성 모두에서 체지방률, 복부비만율, 체질량지수와 같은 체형 및 체성분 관련 지표가 대부분의 모델에서 공통적으로 상위 변수에 포함되었다. 그러나, 상위 순위의 구성과 우선순위에는 성별 차이가 나타났다. 남성에서는 체지방률과 맥박수가 대부분의 모델에서 상위권에 위치하였고, 복부비만율, 수축기혈압, 연령이 함께 상위 변수군을 형성하였다. 반면 여성에서는 체지방률, 복부비만율, 체질량지수가 상위권에 보다 집중되었으며, 연령 또한 다수 모델에서 상위 순위에 포함되었다.

표 5. 남성에서 변수 선택 방법 및 모델별 변수 순위
Table 5. Ranked variables by selection method and model in men

	Statistical model	Machine learning models				
Model	Logistic regression	Random Forest	Logistic regression	AdaBoost	XGBoost	LightGBM
Selection method	Stepwise	FI	FI	RFE	RFE	FI
Rank						
1	PBF	PBF	PBF	Age	PBF	PBF
2	Pulse	Pulse	Pulse	BMI	Pulse	Pulse
3	FHx DM	WHR	FHx DM	Smoking	WHR	WHR
4	SBP	BMI	SBP	PBF	BMI	BMI
5	WHR	SBP	Age	DBP	FHx DM	SBP
6	Age	Age	WHR	WHR	Visceral fat	Muscle
7	HTN hx	Muscle	Hip	Pulse	Muscle	Age
8	Exercise	Visceral fat	BMI	FHx DM	Hip	Hip
9	Smoking	DBP	HTN hx	HTN hx	Education	DBP
10	Walk	Hip	Smoking	SBP	Exercise	Visceral fat

* FI, feature importance; RFE, recursive feature elimination; PBF, percent body fat; DM, diabetes mellitus; FHx, family history; hx, history; HTN, hypertension; WHR, waist-to-hip ratio; SBP, systolic blood pressure; DBP, diastolic blood pressure; circ., circumference.

표 6. 여성에서 변수 선택 방법 및 모델별 변수 순위
Table 6. Ranked variables by selection method and model in women

	Statistical model	Machine learning models				
Model	Logistic regression	Random Forest	Logistic regression	AdaBoost	XGBoost	LightGBM
Selection method	Stepwise	FI	RFE	RFE	RFE	FI
Rank						
1	BMI	PBF	WHR	Age	WHR	PBF
2	WHR	BMI	FHx DM	BMI	SBP	WHR
3	Pulse	WHR	Age	Muscle	FHx DM	BMI
4	Age	Visceral fat	Pulse	PBF	Age	SBP
5	FHx DM	SBP	BMI	SBP	BMI	Pulse
6	SBP	Age	PBF	WHR	PBF	Age
7	HTN hx	Pulse	SBP	Hip circ.	Visceral fat	Muscle
8	Walk	Muscle	Muscle	FHx DM	Education	Hip circ.
9	-	Hip	HTN hx	HTN hx	HTN hx	DBP
10	-	DBP	Hip circ.	Pulse	Drink	FHx DM

* FI, feature importance; RFE, recursive feature elimination; PBF, percent body fat; DM, diabetes mellitus; FHx, family history; hx, history; HTN, hypertension; WHR, waist-to-hip ratio; SBP, systolic blood pressure; DBP, diastolic blood pressure; circ., circumference.

성별 내에서도 알고리즘 및 변수 선택 방법에 따라 변수 순위의 차이가 확인되었다. 남성에서는 체지방률이 모든 모델에서 공통적으로 상위에 위치하였으나, 맥박수, 혈압, 복부비만 관련 지표(WHR, visceral fat)의 상대적 순위는 모델에 따라 달랐다. 여성에서도 체지방률, 복부비만율, 체질량지수가 반복적으로 선택되었지만, 연령, 당뇨 가족력, 혈압, 근육량, 교육 수준, 음주의 포함 여부와 순위는 모델 및 변수 선택 방법에 따라 차이를 보였다. 이러한 결과는 동일한 성별 내에서도 적용한 알고리즘과 변수 선택 방법에 따라 최종 변수 구성과 순위가 달라지는 양상이 관찰되었다.

3.3.2 모델링 결과

표 7은 성별, 모델 및 변수 선택 방법에 따른 예측 성능 비교 결과를 제시한다. 전체적으로 여성 집단의 예측 성능이 남성 집단보다 높게 나타났다. 남성 집단의 AUC는 0.70 – 0.72 범위였으며, 여성 집단의 AUC는 0.77 – 0.80 범위로 나타났다. 재현율(Recall) 또한 남성은 0.73 – 0.81, 여성은 0.78 – 0.88 범위로 여성 집단에서 전반적으로 높았다.

동일한 모델을 사용하더라도 성별에 따라 예측 성능 차이가 확인되었다. 예를 들어 Random Forest(FI)의 AUC는 남성 0.71, 여성 0.77이었고, AdaBoost(RFE)는 남성 0.72, 여성 0.79를 보였다. XGBoost(RFE)와 LightGBM(FI)에서도 유사한 경향이 나타났으며, 전반적으로 여성 집단에서 더 높은 AUC가 관찰되었다.

성별 내에서도 모델 및 변수 선택 방법에 따른 성능 차이가 나타났다. 남성 집단에서는 AdaBoost (RFE)가 AUC 0.72로 가장 높았고, 나머지 모델은 0.70 – 0.71 수준으로 유사한 성능을 보였다. 여성 집단에서는 기계학습 기반 로지스틱 회귀(RFE)와 통계 기반 로지스틱 회귀(Stepwise)가 AUC 0.80으로 가장 높은 성능을 나타냈고, AdaBoost(RFE), XGBoost(RFE), LightGBM(FI)는 0.79 수준을 보였으며 Random Forest(FI)는 0.77로 나타났다.

Ⅳ. 논의 및 결론

본 연구는 KoGES-HEXA 기반의 비침습적 변수만을 이용한 미진단 당뇨 예측에서 성별을 분리하여 통계 기반 방법과 기계학습 기반 방법의 변수 선택 결과 및 예측 성능을 비교하였다. 그 결과, 성별과 알고리즘 차이라는 두 수준에서 뚜렷한 차이가 확인되었다.

표 7. 성별 및 모델별 예측 성능 비교 결과
Table 7. Comparison of predictive performance by sex and model

Sex	Model	Feature selection method	Optimal threshold	Recall	AUC	Accuracy	F1-Score
Men	Random Forest	Feature importance	0.45	0.81	0.71	0.52	0.13
	Logistic Regression	Feature importance	0.47	0.73	0.71	0.60	0.13
	AdaBoost	Recursive Feature Elimination	0.50	0.75	0.72	0.56	0.13
	XGBoost	Recursive Feature Elimination	0.46	0.75	0.70	0.55	0.13
	LightGBM	Feature importance	0.44	0.77	0.71	0.55	0.13
	Logistic Regression	Stepwise	0.04	0.76	0.71	0.56	0.13
Women	Random Forest	Feature importance	0.40	0.83	0.77	0.61	0.09
	Logistic Regression	Recursive Feature Elimination	0.48	0.82	0.80	0.66	0.10
	AdaBoost	Recursive Feature Elimination	0.50	0.84	0.79	0.61	0.09
	XGBoost	Recursive Feature Elimination	0.46	0.85	0.79	0.62	0.10
	LightGBM	Feature importance	0.66	0.78	0.79	0.66	0.10
	Logistic Regression	Stepwise	0.02	0.88	0.80	0.58	0.10

첫째, 남녀 간 차이 측면에서 남성과 여성 모두 체지방률, 체질량지수, 복부비만율 등 체성분 관련 지표가 반복적으로 상위 변수로 선택되었으나, 남성에서는 맥박수와 혈압 관련 지표가 상대적으로 상위에 자주 포함된 반면 여성에서는 연령, 체질량지수, 복부비만율이 더 안정적으로 상위권을 형성하였다. 이러한 양상은 국내 당뇨 유병률 및 연령대별 분포가 성별에 따라 다르게 나타나는 역학적 배경과도 일치하며[7][8], 본 연구에서 성별 분리 모델링을 적용한 설계의 필요성을 뒷받침한다.

둘째, 동일 성별 내에서도 알고리즘 및 변수 선택 방법에 따라 상위 변수의 포함 여부와 순위가 달라졌다. 이는 통계 기반 로지스틱 회귀와 기계학습 기반 모델이 변수 기여도를 평가하는 기준이 다르기 때문으로 볼 수 있다[1][4][5]. 특히 본 연구에서 로지스틱 회귀는 해석 중심의 통계 기준모델(Stepwise)과 예측 비교를 위한 기계학습 기반 분류 모델로 모두 사용되었는데, 동일한 알고리즘이라도 변수 선택 프레임과 목적이 달라질 때 최종 변수 조합과 성능이 달라질 수 있음을 확인하였다. 예측 성능 측면에서는 여성 집단의 AUC(0.77–0.80)가 남성 집단(0.70–0.72)보다 전반적으로 높았고, 같은 모델을 사용하더라도 성별에 따라 AUC와 재현율이 다르게 나타났다. 여성에서는 로지스틱 회귀가 가장 높은 AUC를 보인 반면, 남성에서는 AdaBoost가 가장 높은 AUC를 보였다. 이는 머신러닝이 로지스틱 회귀보다 항상 우수하지 않다는 기존 문헌의 보고[6] 및 국민건강영양조사 기반 연구에서 통계모델과 머신러닝 간 성능 차이가 제한적으로 보고된 선행연구[9]와 일치한다. 동시에 이러한 결과는 성별 분리 여부와 변수 선택 방법이 모델 성능 비교에 실제적인 영향을 줄 수 있음을 추가로 보여준다. 또한 비침습적 변수만으로 미진단 당뇨를 선별하고자 한 접근은 공복 유지와 채혈 부담이 있는 진단 전 단계에서 활용 가능성을 갖는다는 점에서 의의가 있으며[9][10], 본 연구는 이를 KoGES의 대규모 코호트 자료에 적용하여 성별별 변수 선택 패턴과 성능 차이를 함께 제시하였다는 점에서 기존 여성 단일 집단 기반 연구의 범위를 확장한다[10]. 특히 본 연구는 기존 연구에서 성별을 단순 보정 변수로 처리한 것과 달리, 성별을 분리하여 변수 선택 결과 자체의 차이를 비교하였다. 또한 동일한 변수 후보군을 기반으로 통계 기반 방법과 머신러닝 기법을 함께 적용하여 알고리즘에 따라 선택되는 변수 구성과 순위가 어떻게 달라지는지를 확인하였다. 나아가 이러한 변수 선택 결과를 예측 성능과 함께 해석함으로써 성별에 따른 예측모델 구성의 차이를 보다 구체적으로 제시하였다. 이를 통해 본 연구는 단순한 성능 비교를 넘어서 변수 선택 구조와 해석 측면에서의 차이를 함께 보여주었다는 의의를 갖는다.

한편, 본 연구는 당뇨 비율이 남성 4.4%, 여성 2.4%로 낮아 클래스 불균형의 영향을 받을 수 있고, 실제로 Accuracy에 비해 F1-score가 낮게 나타난 점은 이러한 자료 특성을 반영한 결과로 해석된다. 또한 Youden index 기반 임계값 최적화는 민감도와 특이도의 균형을 고려한 방법으로, 클래스 불균형 환경에서는 재현율을 높이는 대신 precision 및 F1-score가 낮아지고 위양성이 증가할 가능성이 있다. 다만 본 연구는 채혈 전 단계에서 미진단 당뇨 고위험군을 우선 선별하기 위한 비침습적 1차 필터를 목표로 하였으므로, 환자 누락을 줄이는 재현율 확보를 우선하는 상황을 전제로 분석하였다. 따라서 실제 적용에서는 목적(1차 선별 vs 확진 보조)에 따라 precision 및 양성예측도(PPV) 등 지표를 함께 고려할 필요가 있다. 실제 집단의 발생 비율을 유지한 상태에서 성별 간 예측 성능과 변수 선택 결과를 비교하는 것이 주요 목적이기 때문에 SMOTE 나 ADASYN과 같은 재표본화 기법을 적용하지 않았다. 다만 향후 연구에서는 균형조정 기반 민감도 분석을 수행하여 예측 성능의 안정성과 잠재적 모델 편향 여부를 평가할 것이다. 또한 본 연구는 예측변수의 결측치가 존재하는 대상자를 제외하는 complete-case 분석을 적용하였으며, 이 과정에서 전체 표본의 상당수가 제외되어 표본 대표성이 저하되거나 선택 편향이 발생했을 가능성을 배제할 수 없다. 다만 본 연구의 주요 목적이 성별 분리 환경에서 알고리즘 및 변수 선택 방법에 따른 결과를 동일한 입력정보 조건에서 비교하는 데 있었으므로, 결측치 대체에 따른 추가 가정과 불확실성을 최소화하고 비교의 일관성을 확보하기 위해 complete-case 분석을 적용하였다. 그리고, 외부 검증

데이터를 활용하지 못해 일반화 가능성 평가가 제한되고, 기반조사 시점 변수와 검사값으로 결과를 정의한 설계 특성상 역인과 가능성을 줄이려는 조치를 취했더라도 완전히 배제할 수는 없다는 제한이 있다. 본 연구는 주로 AUC와 분류 지표 중심으로 성능을 비교하였으며, 예측모델 평가에서 중요한 보정(Calibration) 측면이 포함되지 않았다는 점도 해석 시 고려가 필요하다[6]. 또한 질환 영역과 자료 구조에 따라 통계모델과 머신러닝의 상대적 성능이 달라질 수 있다는 보고들을 감안할 때[14][15], 후속 연구에서는 동일한 성별 분리 전략을 다른 질환 예측 문제와 외부 코호트에 적용한 추가 비교가 요구된다.

본 연구에서는 성별에 따라 최종 변수의 구성과 예측모델의 성능이 실질적으로 달라질 수 있음을 확인하였으며, 알고리즘 선택 역시 모델 구성과 성능에 중요한 영향을 미칠 수 있음을 보여주었다. 이는 질병 예측모델 개발 시 성별을 단순 공변량으로 처리하는 접근을 넘어, 성별 맞춤형 예측모델 설계를 고려할 필요가 있음을 시사한다. 또한 통계기반 및 기계학습 방법의 특성과 변수 선택 기준을 충분히 이해하고, 최종 변수의 구성과 해석을 병행하는 개발 전략이 중요함을 강조한다.

References

[1] H. Park, "An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain", Journal of Korean Academy of Nursing, Vol. 43, No. 2, pp. 154-164, Apr. 2013. https://doi.org/10.4040/jkan.2013.43.2.154.

[2] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine", The New England Journal of Medicine, Vol. 380, No. 14, pp. 1347-1358, Apr. 2019. https://doi.org/10.1056/NEJMra1814259.

[3] S. Rose, "Intersections of machine learning and epidemiological methods for health services research", International Journal of Epidemiology, Vol. 49, No. 6, pp. 1763-1770, Dec. 2020. https://doi.org/10.1093/ije/dyaa035.

[4] D. Bzdok, N. Altman, and M. Krzywinski, "Statistics versus machine learning", Nature Methods, Vol. 15, No. 4, pp. 233-234, Apr. 2018. https://doi.org/10.1038/nmeth.4642.

[5] L. Ryu and K. Han, "Machine Learning vs. Statistical Model for Prediction Modelling: Application in Medical Imaging Research", Journal of the Korean Society of Radiology, Vol. 83, No. 6, pp. 1219-1228, Dec. 2022. https://doi.org/10.3348/jksr.2022.0111.

[6] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. V. Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models", Journal of Clinical Epidemiology, Vol. 110, pp. 12-22, Jun. 2019. https://doi.org/10.1016/j.jclinepi.2019.02.004.

[7] Ministry of Health and Welfare (Korea) and Korea Disease Control and Prevention Agency, Korea Health Statistics 2022: Korea National Health and Nutrition Examination Survey (KNHANES IX-1), https://knhanes.kdca.go.kr/knhanes/sub04/sub04_04_01.do. [accessed: Jan. 27, 2026].

[8] H.-S. Kwon, "Prevalence and treatment status of diabetes mellitus in Korea", Journal of the Korean Medical Association, Vol. 66, No. 7, pp. 404-407, Jul. 2023. https://doi.org/10.5124/jkma.2023.66.7.404.

[9] S. G. Choi, et al., "Comparisons of the prediction models for undiagnosed diabetes between machine learning versus traditional statistical methods", Scientific Reports, Vol. 13, No. 1, Art. No. 13101, Aug. 2023. https://doi.org/10.1038/s41598-023-40170-0.

[10] M. H. Yim, Y. J. Jeon, and H. Kim, "Classification of Diabetes and Impaired Fasting Glucose Using Noninvasive Factors Based on Machine Learning Approaches in Korean Middle-Aged Women", Journal of the Korea Institute of Information Technology, Vol. 21, No. 8, pp. 175-184, Aug. 2023. https://doi.org/10.14801/jkiit.2023.21.8.175

[11] Y. Kim, B. G. Han, and KoGES Group, "Cohort Profile: The Korean Genome and Epidemiology Study (KoGES) Consortium", International Journal of Epidemiology, Vol. 46, No. 2, p. e20, Apr. 2017. https://doi.org/10.1093/ije/dyv316.

[12] F. C. Bull, et al., "World Health Organization 2020 guidelines on physical activity and sedentary behaviour", British Journal of Sports Medicine, Vol. 54, No. 24, pp. 1451-1462, Dec. 2020. https://doi.org/10.1136/bjsports-2020-102955.

[13] C. L. A. Navarro, et al., "Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review", BMJ, Vol. 375, Art. No. n2281, Oct. 2021. https://doi.org/10.1136/bmj.n2281.

[14] H. Sufriyana, et al., "Comparison of Multivariable Logistic Regression and Other Machine Learning Algorithms for Prognostic Prediction Studies in Pregnancy Care: Systematic Review and Meta-Analysis", JMIR Medical Informatics, Vol. 8, No. 11, Art. No. e16503, Nov. 2020. https://doi.org/10.2196/16503.

[15] Z. Sun, et al., "Comparing Machine Learning Models and Statistical Models for Predicting Heart Failure Events: A Systematic Review and Meta-Analysis", Frontiers in Cardiovascular Medicine, Vol. 9, Art. No. 812276, Apr. 2022. https://doi.org/10.3389/fcvm.2022.812276.

저자소개

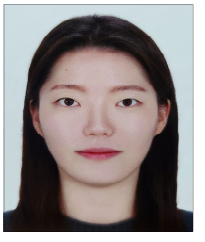
김 정 희 (Jeong-Hee Kim)

2025년 8월 : 을지대학교 빅데이터의료융합학과(공학사)
관심분야 : 빅데이터, 머신러닝, 데이터분석

김 유 빈 (Yu-Bin Kim)

2026년 2월 : 을지대학교 빅데이터의료융합학과(공학사)
관심분야 : 빅데이터 분석, 머신러닝, 데이터 시각화

서 주 희 (Ju-Hee Seo)



2021년 3월 ~ 현재 : 을지대학교 빅데이터의료융합학과 학사과정
관심분야 : 질병예측, 머신러닝, 빅데이터 분석

정 재 효 (Jaehyo Jung)

2015년 9월 ~ 2018년 12월 : 조선대학교 IT융합신기술연구센터 연구원
2019년 2월 : 조선대학교 IT융합학과(공학박사)
2019년 3월 ~ 2025년 2월 : 조선대학교 T융합신기술연구센터 연구교수
2025년 3월 ~ 현재 : 을지대학교 첨단학부 빅데이터인공지능전공 조교수
관심분야 : 생체신호 취득 시스템, 질병예측, 딥러닝, 패턴인식

박 주 용 (JooYong Park)

2015년 2월 : 성균관대학교 유전공학과(이학사)
2022년 2월 : 서울대학교 의과학과(의학박사)
2020년 2월 ~ 2020년 9월 : 서울대학교 의학연구원 연수연구원
2022년 10월 ~ 현재 : 을지대학교 빅데이터의료융합학과 조교수
2025년 3월 ~ 현재 : 을지대학교 첨단학부 빅데이터인공지능전공 조교수
관심분야 : 의료빅데이터, 질병예측, 정밀의료