Regular paper | Journal of the KIECS. pp. 999-1008, vol. 18, no. 5, Oct. 31. 2023, t. 121, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2023.18.5.999

빅데이터 기반 2형 당뇨 예측 알고리즘 개발

심 현* · 김현욱

Development of Type 2 Prediction Prediction Based on Big Data

Hyun Sim* · HyunWook Kim

요 약

당뇨병과 같은 만성 질환의 조기 예측은 중요한 이슈이며, 그중에서도 당뇨 예측의 정확도 향상은 매우 중요하다. 당뇨 예측을 위한 다양한 기계 학습 및 딥 러닝 기반 방법론을 도입하고 있으나, 이러한 기술들은 다른 방법론보다 더 우수한 성능을 위해 대량의 데이터를 필요로 하며, 복잡한 데이터 모델 때문에 학습 비용이 높다. 본 연구에서는 pima 데이터셋과 k-fold 교차 검증을 사용한 DNN이 당뇨 진단 모델의 효율성을 감소시킨다는 주장을 검증하고자 한다. 의사 결정 트리, SVM, 랜덤 포레스트, 로지스틱 회귀, KNN 및 다양한 앙상블 기법과 같은 기계 학습 분류 방법을 사용하여 어떤 알고리즘이 최상의 예측 결과를 내는지 결정하였다. 모든 분류 모델에 대한 훈련 및 테스트 후 제안된 시스템은 ADASYN 방법과 함께 XGBoost 분류기에서 최상의 결과를 제공하였으며, 정확도는 81%, F1 계수는 0.81, AUC는 0.84였다. 또한 도메인 적응 방법이 제안된 시스템의 다양성을 보여주기 위해 구현되었다. LIME 및 SHAP 프레임워크를 사용한 설명 가능한 AI 접근 방식이 모델이 최종 결과를 어떻게 예측하는지 이해하기 위해 구현되었다.

ABSTRACT

Early prediction of chronic diseases such as diabetes is an important issue, and improving the accuracy of diabetes prediction is especially important. Various machine learning and deep learning-based methodologies are being introduced for diabetes prediction, but these technologies require large amounts of data for better performance than other methodologies, and the learning cost is high due to complex data models. In this study, we aim to verify the claim that DNN using the pima dataset and k-fold cross-validation reduces the efficiency of diabetes diagnosis models. Machine learning classification methods such as decision trees, SVM, random forests, logistic regression, KNN, and various ensemble techniques were used to determine which algorithm produces the best prediction results. After training and testing all classification models, the proposed system provided the best results on XGBoost classifier with ADASYN method, with accuracy of 81%, F1 coefficient of 0.81, and AUC of 0.84. Additionally, a domain adaptation method was implemented to demonstrate the versatility of the proposed system. An explainable AI approach using the LIME and SHAP frameworks was implemented to understand how the model predicts the final outcome..

키워드

AI, Diabetes prediction, Deep Learning, Machine Learning, Over Sampling
인공지능, 당뇨예측, 딥러닝, 머신러닝, 오버샘플링

* 순천대학교 부교수(simhyun@scnu.ac.krl)
* 교신저자 : 순천대학교 스마트농업전공
· 접 수 일 : 2023. 08. 27
· 수정완료일 : 2023. 09. 19
· 게재확정일 : 2023. 10. 17
· Received : Aug. 27, 2023, Revised : Sep. 19 2023, Accepted : Oct. 17, 2023
· Corresponding Author : Hyun Sim
Dept. Smart Agriculture, Sunchon National University
Email : simhyun@scnu.ac.kr

Ⅰ. 서 론

전 세계적으로 당뇨병은 5억 3,700만 명에게 영향을 미쳐, 가장 치명적이면서 가장 흔한 비전염성 질병이다. 과체중, 콜레스테롤 수준, 가족력, 신체 활동 부족, 나쁜 식습관 등 많은 요인이 당뇨병에 영향을 받게 된다. 오랜 시간 동안 당뇨병을 앓은 사람들은 심장 장애, 신장 질환, 신경 손상, 당뇨 망막병증 등 여러 합병증을 얻을 수 있다. 그러나 조기에 예측되면 그 위험을 줄일 수 있다. 2형당뇨의 유병률은 다양한 사회적 요인으로 인해 빠르게 증가하고 있으며, 사회와 개인에게 부담도 증가하고 있다. 최근에는 복합 만성 질환이 증가하는 경향이 있다. 2형 당뇨병은 생활 습관과 밀접한 관련이 있으며, 여러 생활 습관 요인의 조합에 의해 발생하는 것으로 알려져 있다. 2형 당뇨병의 효과적인 관리는 예방을 중심으로 하는 변화하는 건강관리 패러다임에 중점을 둘 필요가 있다.

인공 지능과 제4차 산업혁명의 발전으로 다양한 질병, 특히 2형당뇨의 진단 및 예측이 현실화 되었다. 이러한 진단 및 예측의 효율성은 빅데이터와 AI 학습 모델의 문제점을 극복함으로써 높아진다. 의료 데이터에서는 누락 값, 이상치, 클래스 불균형 등의 문제가 발생할 수 있으며, 이는 분석의 편향성과 성능 저하를 초래한다. 2형 당뇨의 유병률은 증가하는 추세로, 생활 습관과 밀접한 연관이 있다. 따라서 예방 중심의 관리 전략이 필요하다. 빅 데이터와 인공 지능은 건강관리 분야에서 중요한 역할을 하며, 전 세계적인 연구 동향에 따라 한국에서도 이에 대한 연구가 확대되고 있다. 본 연구에서는 상호 정보 특성 선택 알고리즘을 적용하였다. 사설 데이터셋의 인슐린 특성을 예측하기 위해 극단적인 경사 부스팅을 사용한 반지도 학습 모델이 활용되었다. SMOTE(: Synthetic Minority Over-sampling Technique)와 ADASYN(: Adaptive Synthetic Sampling Approach) 방법이 클래스 불균형 문제를 관리하기 위해 사용되었다. 의사 결정 트리, SVM(: Support Vector Machines), 랜덤 포레스트, 로지스틱 회귀, KNN(: K-Nearest Neighbors) 및 다양한 앙상블 기법과 같은 기계 학습 분류 방법을 사용하여 어떤 알고리즘이 최상의 예측 결과를 내는지 결정하였다. 모든 분류 모델에 대한 훈련 및 테스트 후 제안된 시스템은 ADASYN 방법과 함께 XGBoost 분류기에서 최상의 결과를 제공하였으며, 정확도는 81%, F1 계수는 0.81, AUC(: Area Under the Curve)는 0.84였다. 또한 도메인 적응 방법이 제안된 시스템의 다양성을 보여주기 위해 구현되었다. LIME(: Local Interpretable Model-Agnostic Explanation) 및 SHAP(: Shapley Additive exPlanations) 프레임워크를 사용한 설명 가능한 AI 접근 방식이 모델이 최종 결과를 어떻게 예측하는지 이해하기 위해 구현되었다.

이 연구는 기계 학습을 통한 당뇨병 예측을 구현하는 것에 관하여 작성하였다. 주요 성과는 다음과 같다. SMOTE와 ADASYN 기법은 클래스 불균형 문제를 최소화하기 위해 구현되었다. 이 작업에서는 하이퍼파라미터 튜닝도 일부 수행하였다. SHAP과 LIME 라이브러리를 사용한 설명 가능한 AI 기법이 구현되어 모델이 어떻게 결정을 예측하는지 이해할 수 있다. 이 접근법은 어떤 특성이 예측 측면에서 가장 중요한 역할을 하는지 해석하는 데 도움이 된다. 이 연구 개발을 통해 최종적으로 선정된 최적의 모델로 웹과 어플리케이션을 개발하여 실시간 데이터로 즉시 예측을 할 수 있을 것으로 예상한다. 본 연구에서 중요한 점은 기계 학습 및 앙상블 기법을 사용하였다는 것이다. 본 논문의 구성은 다음과 같다. 2장에서 관련 연구를 분석한다. 3장에서 본 연구에서 구현한 시스템을 이미지 및 플로우차트와 함께 설명한다. 연구의 최종 결과는 4장에서 제시하였으며, 5장에서는 본 연구 결과를 요약하면서 한계점을 살펴보고 추후 연구 방안에 대해 기술하였다.

Ⅱ. 관련 연구

2.1 Deep Learning 접근법

비지도 학습 접근법인 딥 뉴럴 네트워크(DNN)를 사용한 피마 인디언 당뇨병 데이터셋에 대한 예측 모델은 98.16%의 정확도를 달성하였으며[1-2], 다섯 번과 열 번의 교차 검증으로 속성을 훈련시키는 딥 뉴럴 네트워크를 사용한 당뇨병 진단 전략으로 얻어진 정확도는 98.35%로 나타났다[3]. 이진 교차 엔트로피 손실 함수와 여러 매개변수를 사용하는 모델을 제안하였다. PIMA 데이터베이스에서 딥 뉴럴 네트워크를 사용하는 예측 모델은 약 99.41%의 성능을 보였으며

[4], Bala 외는 VRC(심박수 변동)를 데이터로 사용하고 CNN-LSTM(LSTM = Long Short Term Memory)을 사용하여 이상을 자동으로 감지하였다. 5-중첩 교차 검증과 CNN을 사용하였고, 정확도는 93.6%를 달성하였으며, CNN-LSTM 조합은 최대 95.1%의 정확도를 보였다[5]. 또한 합성곱 장기 기억(CLSTM) 기반의 접근법이 제안되었으며, 당뇨병 분류 모델이 개발되었고 Pima Indians Diabetes Database (PIDD)에서 기존 방법론과 비교되었다. CLSTM 모델에 의해 얻어진 결과는 다른 방법론보다 높은 96.8%이고[6], PIMA를 사용하여 다양한 기계 학습 알고리즘을 사용한 당뇨병 예측 방법론을 제시한 DL과 DT 데이터셋은 98.07%의 높은 정확도를 제공하였다[7].

2.2 Machine Learning 접근법

딥 러닝이 나타나기 이전에 기계 학습 알고리즘이 큰 역할을 하는 주제로 당뇨병의 조기 진단이 있었다.

Naz 외는 기계 학습(ML) 알고리즘과 신경망(NN) 방법이 사용하였는데 이 연구진은 로지스틱 회귀(LR)와 서포트 벡터 머신(SVM) 모델이 88.6%의 정확도로 당뇨병 예측에 잘 작동한다는 것을 발견했다. 여섯 가지 학습 기반 분류 방법이 온라인 및 오프라인 설문지를 통해 수집된 데이터셋에 적용되었으며, 동일한 알고리즘이 PIMA 데이터베이스에도 적용되었다. 실험 결과, 데이터셋 pima에 대한 랜덤 포레스트의 정확도는 94.10%로, 다른 것들 중에서 가장 높았다[25]. 많은 데이터셋에서 많은 기계 학습 알고리즘을 훈련한 SVM이 98.6%의 정확도로 다른 알고리즘을 능가하는 것으로 나타났다[8]. Chowdary 외 연구자들은 인도 피마 환자의 의료 기록에서 당뇨병을 효과적으로 예측하기 위한 새로운 접근법을 제안하였다. 데이터 마이닝 절차의 정확도를 높이기 위해 수정된 J48 분류기가 사용되었다. WEKA 데이터 마이닝 도구가 MATLAB의 API로 사용되어 수정된 J-48 분류기를 생성하였다. 실험 결과는 기존의 J-48 알고리즘보다 크게 개선되었다는 것을 보여주었다. 제안된 알고리즘이 99.87%의 정확도를 달성할 수 있다는 것이 증명되었다[9]. Mat Jizat 외는 주성분 분석 방법을 통해 중요한 속성 선택이 이루어졌다. 그들의 연구 결과는 당뇨병과 체질량 지수(BMI) 및 Apriori 방법을 통해 추출된 포도당 수준과의 강력한 연관성을 나타냈다. 인공 신경망(ANN), 랜덤 포레스트(RF) 및 K-평균 군집 기법이 당뇨병 예측을 위해 구현되었다. ANN 기법은 75.7%의 최고 정확도를 제공하였으며, 치료 결정을 돕기 위해 의료 전문가에게 유용할 수 있다[10].

III. 당뇨병 자동 예측 시스템

자동 당뇨병 예측 시스템을 설계하기 위한 다양한 기계 학습 기법의 작동 절차와 구현은 다음과 같다. 데이터셋을 수집하고 데이터셋에서 필요한 차이를 제거하기 위해 데이터 전처리를 수행했다. 예를 들면, null 값, 공백값 등을 평균 값으로 대체하거나 불균형 클래스 문제를 처리하는 것 등이다. 그 다음, 홀드아웃 valid 기법을 사용하여 데이터셋을 학습 세트와 테스트 세트로 분리했다. 다음으로, 이 데이터셋에 가장 적합한 분류 알고리즘을 찾기 위해 다양한 분류 알고리즘이 적용되었다.

3.1 데이터셋 수집

Pima Indian 데이터셋은 기계학습분류를 위해 공개적으로 사용 가능한 오픈 소스 데이터셋[11]이다. 이 데이터셋은 768명의 환자 데이터를 포함하고 있으며, 그 중 268명이 당뇨병을 발병했다. 그림 1은 Pima Indian 데이터셋에서 당뇨병을 가진 사람들의 비율을 보여준다. 표 1에 오픈 소스 Pima Indian 데이터셋의 특징을 정리하였다.

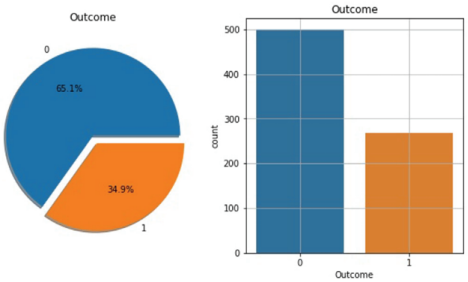

그림 1. 데이터 세트에서 당뇨병을 앓고 있는 사람의 비율
Fig. 1 Proportion of people with diabetes in the data set

표 1. 피마 인디언 데이터 세트의 특징
Table 1. Features of the Pima Indian Data Set

Pregnancies	Skin thickness	Diabetes pedigree function
Glucose	Insulin	Age
Blood pressure	BMI	

3.2 데이터셋 처리

데이터를 처리하며 병합된 데이터셋에서 몇몇 특이한 0 값들이 발견되었다. 예를 들어, 피부 두께와 체질량 지수(BMI)는 0이 될 수 없다. 0 값은 해당 평균값으로 대체되었다. 훈련 및 테스트 데이터셋은 holdout 검증 기법을 사용하여 분리되었으며, 80%는 훈련 데이터, 20%는 테스트 데이터로 사용되었다.

그림 2는 이 데이터셋의 다양한 특징들, 즉, 각 속성의 중요도를 보여준다. 예를 들어, 이 그림에 따르면, 당뇨 유전 기능은 이 상호 정보량 기법에 따라 덜 중요해 보인다.

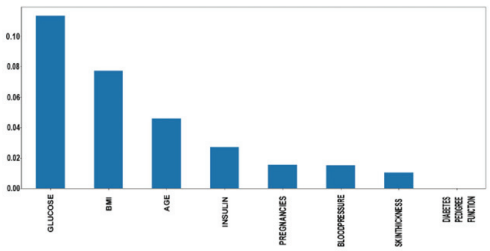

그림 2. 항목별 특징의 중요도
Fig. 2 Importance of characteristics for each item

이 연구에서는 공개 소스인 Pima Indian 데이터셋과 사적인 RTML(: Real-Time Machine Learning) 데이터셋을 결합하여 사용했다. 반지도 학습 방식을 사용하여 예측하였으며 수집된 데이터셋을 Pima Indian 데이터셋과 병합하기 전에, 극단적 경사 부스팅 기법(XGB regressor)을 사용하여 모델을 생성했다. 다양한 회귀 및 앙상블 학습 기법은 누락 된 값 예측에 성공적으로 사용되었다. RTML 데이터셋의 인슐린 특징을 Pima Indian 데이터셋에서 예측하기 위한 최적의 회귀 기법을 선택하면서 광범위한 조사가 수행되었다. RTML 데이터셋에 실제 인슐린 값이 없었기 때문에, Pima Indian 데이터셋을 초기에 사용하여 최상의 회귀 모델을 선택했다. 처음에 Pima Indian 데이터셋은 8:2의 비율로 나누어졌고, 세 가지 지도 학습 회귀 모델, 즉 극단적 경사 부스팅 기법(XGB), 서포트 벡터 회귀(SVR), 그리고 가우시안 프로세스 회귀(GPR)가 사용되어 Pima Indian 데이터셋의 검증 샘플의 선택된 결과, 즉 인슐린을 예측하기 위해 식(1)이 적용되었다.

$$SE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}} \quad \cdots (1)$$

아래 표2에 따르면 XGB 기법은 Pima Indian 데이터셋에서 인슐린에 대한 최소 RMSE(: Root Mean Square Error)를 보여준다. 따라서 이 모델은 Pima Indian 데이터셋에서 수집된 RTML 데이터셋의 누락된 인슐린 열을 예측하는 데 사용되었다. RTML 데이터셋에서 인슐린을 예측하는 작업 단계는 그림 4에서 설명되어 있다. 반지도형 학습 후, 인슐린 특성을 예측하고 RTML 데이터셋을 Pima Indian 데이터셋과 병합했다. 병합된 데이터셋은 모든 특성을 포함하고 있으며, 상호 정보에 따라 가장 중요하지 않은 특성인 당뇨병 유전자 함수를 제외하고 877개의 데이터를 포함하고 있다.

표 2. 피마 인디언 데이터 세트의 다양한 회귀 모델 RMSE
Table 2. RMSE of various regression models on the Pima Indian dataset

Regression model	RMSE
XGB	0.36
SVR	0.45
GPR	0.43

3.3 기계학습 분류

당뇨 예측 알고리즘을 구현하기 위해 다양한 기계학습 및 앙상블 기법을 사용했다. GridSearchCV 프레임워크는 모든 기계 학습 모델에 대한 다양한 하이퍼파라미터의 최적 값을 찾기 위해 사용하였으며 이를 통해 과적합을 방지하려고 했다.

• 결정 트리: 결정 트리는 규칙의 집합으로 제공되는 학습 함수를 나타낸다. 결정 트리 학습 기법은 이산 값 목표 함수를 근사화하는 방법을 수행한다. 지니 계수나 엔트로피는 정보 획득을 결정하는 데 사용되며, 이 계수를 기반으로 각 노드가 선택된다. 이 계수는 다음과 같이 표현된다.

$$Gini_i = 1 - \sum_{k=1}^{n} (P_{i,k})^2 \quad \cdots (2)$$

$$Entropy = \sum_{i=1}^{n} - p_i \log_2 p_i \quad \cdots (3)$$

식(2)와 (3)에서 n은 고유한 클래스 값의 수를 나타낸다. 이 프로젝트에서 사용된 데이터셋에서 GridSearchCV 하이퍼파라미터 조정을 사용하여 maxdepth=2, minimumsamplesleaf=50 및 'Gini' 불순도 지표가 잘 작동함을 확인했다.

• KNN 분류기: K개의 가장 가까운 분류기를 사용하여 이산 값 함수를 근사화할 수 있다. 분류를 위해 사용 가능한 훈련 점들로 평면을 생성하고 쿼리와 훈련된 점들 간의 거리를 계산한다. K(데이터셋에 따라 다름) 개의 이웃을 결정하고 다수결로 분류한다. 이 프로젝트에서는 2진 분류를 위해 K=5를 사용했다.

• 랜덤 포레스트: 랜덤 포레스트는 여러 결정 트리의 예측을 평균화하는 기계 학습 시스템이다. 결과적으로 랜덤 포레스트는 앙상블 학습 모델로 간주될 수 있다. 이 프로젝트에서는 하이퍼파라미터 튜닝을 통해 estimator=400, minimumsamplesleaf=5, 'Gini' 불순도 지표를 적용한 랜덤 포레스트를 사용했다.

• 서포트 벡터 머신(SVM): SVM은 최적의 초평면을 선택하여 지도 분류를 수행한다. 이 연구에서는 훈련 세트에서 다양한 SVM 커널을 실험했다. 최종적으로 linear 커널, 파라미터 C=10 및 gamma=1을 가진 SVM이 이 데이터셋에서 최상의 결과를 생성함을 발견했다.

• 로지스틱 회귀: 로지스틱 회귀는 2진 클래스를 예측하는 데 사용할 수 있다. 결과를 예측하기 위해 'S' 형 함수를 적합한다. 하이퍼파라미터 최적화 기법은 로지스틱 회귀 모델의 수렴을 위한 최대 반복 횟수를 150으로 결정했다.

• AdaBoost: AdaBoost는 앙상블 기법이다. 이 분류기는 원래의 데이터셋에서 초기 작업을 수행한 다음 동일한 데이터셋에 분류기의 반복된 사본을 적합한다. 이 프레임워크는 연속적인 분류기가 어려운 상황에 더 집중하도록 잘못 분류된 인스턴스의 가중치를 조정한다. 본 연구에서는 AdaBoost를 estimator=50 및 학습률 0.10으로 적용했다.

• XGBoost: XGBoost는 결정 트리를 기반으로 한 앙상블 기계 학습 기법이며, 그래디언트 부스팅 접근법을 사용한다.

• Voting classifier: 분류를 개선하기 위해 투표를 사용하는 앙상블 기법이다. 이 연구에서는 'soft' 투표 하이퍼파라미터를 사용하여 각 분류기가 예측한 다수의 클래스를 선택하는 투표 분류기를 구현했다.

• Bagging: Bagging은 원래 데이터셋의 무작위 부분 집합에 기본 분류기를 적용하는 앙상블 분류방식이다. 그리고 개별 예측을 집계하여 최종 분류를 생성한다. 구현된 배깅 분류기는 기본 추정량이 500, 최대 샘플 수는 100, 그리고 out-of-bag 점수는 'True'로 설정되어 다양한 하이퍼파라미터로 사용된다.

3.4 당뇨 예측 시스템 결과

제안된 당뇨 예측 시스템의 결과와 내용에 대해 정리하였다. 먼저 다양한 기계 학습 기술의 성능에 대해 논의한다. 정밀도, 재현율, F1 점수, AUC 및 분류 정확도를 사용하여 다양한 ML 모델을 평가하였다. 이러한 지표의 방정식은 다음과 같이 표현된다.

$$\mathrm{Precision} = \frac{TP}{TP + FP} \quad \cdots (4)$$

$$Recall = \frac{TP}{TP + FN} \quad \cdots (5)$$

$$F1\,score = \frac{2 \times Recall \times \mathrm{Precision}}{Recall + \mathrm{Precision}} \quad \cdots (6)$$

여기서 TP는 모델이 긍정적으로 예측하고 결과도 긍정적인 것을 나타낸다. FP는 모델이 긍정적으로 예측은 하지만 결과는 부정적인 것을 나타낸다. TN은 모델이 부정적으로 예측하고 결과도 부정적인 것을 나타낸다. FN은 모델이 부정적으로 예측하지만 결과

는 긍정적인 것을 나타낸다. 이번 연구에서는 모든 기계 학습 모델에 대해 계층화된 8:2 훈련-테스트 분할을 가진 홀드아웃 검증 방식이 사용되었다.

표 3. 데이터 세트의 SMOTE 기법을 사용한 알고리즘의 성능 지표
Table 3. Performance metrics of the algorithm using SMOTE technique on the dataset

Classifier	Precision	Recall	F1 Score	Accuracy	AUC
Logistic regression	0.78	0.77	0.77	77%	0.88
KNN	0.78	0.76	0.76	76%	0.85
Random forest	0.78	0.78	0.78	78%	0.87
Decision tree	0.75	0.73	0.73	73%	0.75
Bagging	0.80	0.79	0.79	79%	0.87
Adaboost	0.79	0.78	0.78	78%	0.85
XGboost	0.78	0.78	0.78	78%	0.84
Voting	0.79	0.79	0.79	79%	0.86
SVM	0.78	0.75	0.76	75%	0.87

표3은 SMOTE 합성 오버샘플링 기술로 병합된 데이터 세트에 대한 다양한 분류기의 성능 지표를 비교한다. 이 표에 따라, 배깅 분류기는 79%의 정확도와 각각 0.79와 0.87의 F1 점수와 AUC로 가장 좋은 전반적인 성능을 보인다.

표 4. 데이터 세트에서 adasyn을 사용하는 다양한 알고리즘의 성능 지표
Table 4. Performance metrics of various algorithms using adasyn on our dataset

Classifier	Precision	Recall	F1 Score	Accuracy	AUC
Logistic regression	0.76	0.75	0.75	75%	0.84
KNN	0.76	0.73	0.73	73%	0.82
Random forest	0.76	0.76	0.76	76%	0.84
Decision tree	0.81	0.72	0.72	72%	0.78
Bagging	0.80	0.79	0.79	79%	0.84
Adaboost	0.75	0.76	0.76	76%	0.84
XGboost	0.81	0.81	0.81	81%	0.84
Voting	0.77	0.77	0.77	77%	0.84
SVM	0.78	0.78	0.77	78%	0.83

표4는 병합된 데이터 세트에서 ADASYN 방법을 사용하여 모든 분류기의 다양한 성능 지표를 보여준다. 표3에 따라 XGBoost 프레임워크는 81%의 정확도와 0.84 AUC로 다른 분류기보다 더 나은 성능을 보인다. 반대로, 의사결정트리 방법은 가장 낮은 정확도와 F1 점수를 보인다.

(1) 도메인 적응 접근법

본 연구에서 적용된 도메인 적응 접근법은 다음과 같은 결과를 보여준다. 기계 학습 모델은 다른 샘플, 즉 원본 및 대상 데이터 세트에서 각각 훈련 및 평가했다. 초기에 당뇨병 예측 모델은 크기가 더 큰 오픈 소스 Pima Indian 데이터 세트에서 훈련하였다. 마지막으로, 이 모델은 훨씬 작은 차원의 사설 RTML 데이터 세트에서 평가된다. 표4는 사설 데이터 세트의 성능 지표를 보여준다. 이 경우에 훈련 데이터 세트에서 XGBoost와 ADASYN 프레임워크가 적용되었다.

표 5. 프라이빗 데이터 세트의 성능 메트릭(도메인 적응기술)
Table 5. Performance metrics for private data sets (domain adaptation technology)

Precision	Recall	F1 Score	Accuracy
0.95	0.96	0.95	96%

(2) ADASYN을 사용한 XGBoost의 혼동 행렬

그림3은 ADASYN을 사용한 XGBoost의 혼동 행렬을 나타낸다. 이 그림에 따르면 XGBoost 기법은 TP = 43 및 TN = 98로 141개의 인스턴스를 올바르게 분류했다.

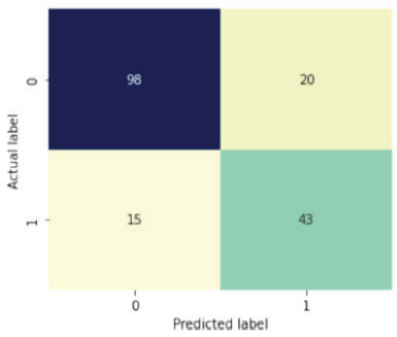

그림 3. ADASYN 기술이 적용된 XGBoost의 매트릭스
Fig. 3 Matrix of XGBoost with ADASYN technology

ADASYN 접근법을 사용한 XGBoost의 ROC(: Receiver Operating Characteristic) 곡선은 그림4에 나타나 있다. 이 그림은 XGBoost의 AUC 값이 0.84임을 보여준다.

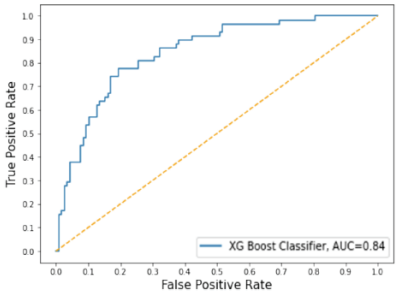

그림 4. ADASYN을 사용한 XGBoost의 ROC 곡선 및 AUC
Fig. 4 ROC curve and AUC for XGBoost with ADASYN

다음으로, SHAP 및 LIME 프레임워크를 사용한 설명 가능한 AI 기법이 구현되어 모델이 어떻게 결정을 예측하는지 이해한다. 그림5는 설명 가능한 AI, SHAP 라이브러리를 사용하여 ADASYN과 함께 사용한 XGBoost의 특징 중요도를 보여준다.

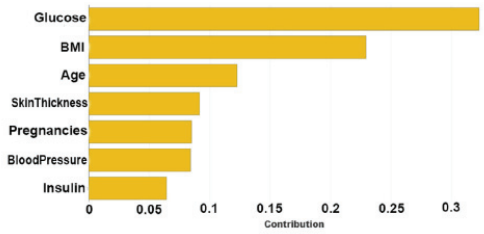

그림 5. XGBoost와 ADASYN기능의 중요도에 대한 AI 해석
Fig. 5 AI analysis of the importance of XGBoost and ADASYN functions

그림 6은 LIME 설명 가능한 AI 방법으로 구현된 XGBoost 모델의 해석을 보여준다. 이 그림에 따르면, 모델은 이 특정한 사람에 대해 당뇨병을 80%의 확신으로 올바르게 예측한다. ML모델은 이 사람이 140.25보다 높은 포도당 수준을 가지고 있으며 6회 이상의 임신을 경험했기 때문에 이 클래스를 예측한다.

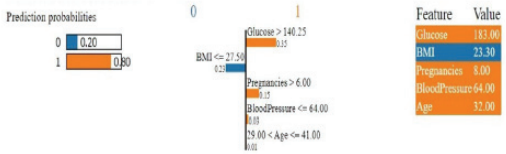

그림 6. LIME AI 예측 해석
Fig. 6 LIME AI Predictive Analysis

마지막으로, 이 연구의 당뇨병 예측 시스템은 ADASYN과 함께 XGBoost 머신러닝 프레임워크를 사용하여 테스트를 진행하였다. 그림 7은 설계된 테스트 애플리케이션을 사용하여 실제 데이터로 즉시 당뇨병 예측을 보여준다.

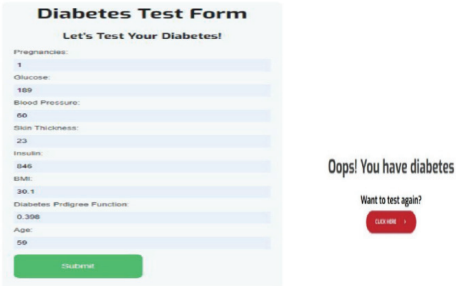

그림 7. LIME AI 예측 해석 테스트 어플리케이션을 통한 당뇨 예측
Fig. 7 Diabetes prediction through LIME AI predictive interpretation test application

마지막으로 Pima Indian 데이터셋의 인슐린 특성을 제거하여 RTML 데이터셋과의 일관성을 유지하는 시나리오를 고려하였다. 표 6은 인슐린 특성을 제거한 후 병합된 데이터셋의 다양한 성능 지표를 보여준다. 이 표에 따르면, 모든 예측 모델의 성능이 저하되었음을 확인할 수 있다.

표 6. 병합된 데이터 세트의 성능 메트릭(인슐린 피마 인디언에서 제거됨)]
Table 6. Performance metrics from merged dataset (removed from Insulin Pima Indian)]

Classifier	Precision	Recall	F1 Score	Accuracy
AdaBoost	0.73	0.71	0.72	72%
Random Forest	0.72	0.70	0.71	71%
XGBoost	0.74	0.73	0.73	74%

Ⅳ. 연구 결과

4.1 알고리즘별 성능검증

표 7에 Pima Indian 데이터셋에 대한 유사한 프로젝트(레퍼런스)와 본 연구의 당뇨병 예측 시스템의 성능 비교를 정리하였다. 표의 메트릭에 따라, 제안된 ADASYN을 사용한 XGBoost 기법은 정확도와 F1 점수 측면에서 대부분의 기존 케이스를 능가하였다. (데이터의 양과 분석 방법에 차이가 있을수 있음.)

표 7. 제안 시스템과 유사한 당뇨 예측 비교
Table 7. Experimental results

Classifier	F1 Score	Accuracy	Other metrics
Deep belief network model	0.81	N/A	Presicion: 0.68 Recall: 1.0
SVM with RBF kernel		82%	
SVM	0.73	75%	Presicion: 0.72 Recall: 0.75
Ensemble(XGBoost)	0.81	88.8%	Presicion: 0.84 Recall: 0.79
Soft voting	0.72	79.1%	Presicion: 0.73 Recall: 0.72
XGBoost with ADASYN	**0.81**	**88.5%**	**Presicion: 0.82 Recall: 0.80**

이 프로젝트의 목적은 기계 학습을 활용해 당뇨를 자동으로 예측하는 것이다. Pima Indian 데이터셋과 캐글, UCI 데이터를 포함한 새로운 RTML 데이터셋을 사용하였다. RTML 데이터셋의 누락된 인슐린 특성 값은 Pima Indian 데이터셋에서 예측하였다. 프로젝트 결과, XGB 회귀 기법은 인슐린을 예측하는 데 가장 낮은 RMS(: Root mean square) 오차를 달성했다는 것을 확인하였다. 상호 정보 기반 특성 선택 알고리즘은 당뇨병 예측에서 글루코스 수준, BMI, 나이, 그리고 인슐린이 가장 중요한 특성임을 나타냈다. SMOTE와 ADASYN 합성 데이터 오버샘플링 및 하이퍼 파라미터 최적화 기법을 적용하였다. ADASYN을 사용한 XGBoost 기법은 최상의 성능을 보였다. LIME 및 SHAP 해석 가능한 AI 프레임워크는 ML 접근법에 의해 제공된 예측을 해석하였다. 이 연구의 한계는 사용된 RTML 데이터셋의 인슐린 특성을 사용할 수 없다는 것이다. XGB 회귀자로부터 얻은 인슐린 예측과 Pima India 데이터셋의 평균 및 중앙값에서 생성된 예측은 분류 정확도에 대해 각각 약 1.33%와 2.33%의 평균 편차를 포함하였다.

4.2 최종 모델

딥 러닝을 사용한 많은 연구가 좋은 결과를 얻었다. 이 연구는 Pima Indian Diabetes Dataset (PIDD), Deep Neural Network(DNN) 및 10 k-fold 교차 검증을 사용하여 모델을 평가하는 데 구현되었다. 분류 작업에서 데이터베이스 내의 클래스 분포는 균형을 이루지 못할 수 있다. k-fold 교차 검증은 pima 데이터셋의 사용 가능한 데이터를 더 잘 활용하는 데 도움을 주었다. k-fold 교차 검증 절차가 끝나면 k개의 성능 점수가 얻어진다. 이 k개의 성능 점수의 평균과 표준 편차를 계산하여 검증 성능의 편향과 분산을 추정할 수 있다. 결과적으로 10 fold 교차 검증은 DNN과 PIDD에 구축된 모델의 성능을 저하시킬 수 있다는 것을 보여주었다. 이 연구에서는 DNN을 사용한 당뇨병 예측에 대한 이전 연구의 비교 분석을 제공하여 10 k-fold 교차 검증과 DNN이 당뇨병 예측 모델의 성능을 어떻게 저하시키는지 보여준다.

Ⅴ. 결론

본 연구에서는 당뇨병은 세계적으로 가속화되고 있는 중대한 건강 위험 요소로, 이로 인해 수명과 삶의 질이 저하되고 있다. 특히 초기 단계에서의 진단은 생명을 구하고, 장기적인 합병증을 예방하는데 결정적인 중요성을 갖는다. 그러나 현실에서는 많은 국가, 특히 개발 중이거나 후진국에서 이러한 초기 단계 예측의 중요성을 충분히 인식하지 못하고 있다.

기계 학습 및 딥 러닝은 이러한 문제의 해결 방안 중 하나로 떠오르고 있다. 본 보고서에서 소개된 연구는 Pima Indian 데이터셋을 중심으로 다양한 기계 학습 모델을 활용한 당뇨병 예측 시스템을 제안하고 평가하였다. 특히, 데이터 불균형 문제를 해결하기 위한

SMOTE와 ADASYN 전처리 기법의 적용, 그리고 여러 머신러닝 및 앙상블 기법의 성능을 비교 분석하였다. 이 중에서도 XGBoost 분류기는 81%의 높은 정확도를 보였고, 이를 바탕으로 도메인 적응 기법의 활용 가능성을 제시하였다.

그러나 이러한 높은 성능에도 불구하고 k-fold 교차 검증을 통한 결과는 모델의 효율성 감소를 시사하였다. 이는 향후 연구의 방향성을 제시하며, 더 큰 환자 집단 및 추가 데이터 확보, 그리고 다양한 기술 결합을 통한 최적화 방안을 탐색할 필요성을 강조한다.

기계 학습과 딥 러닝을 활용한 당뇨병 예측은 매우 유망한 분야로, 이를 통해 초기 단계에서의 당뇨병 진단 및 예방이 가능하게 된다. 본 연구의 실질적인 효용은 그것이 어떻게 현실 세계의 문제, 특히 초기 단계에서의 당뇨병 진단에 어떻게 적용될 수 있는지에 대한 근거에 크게 의존하게 된다. 이후의 연구는 다양한 인구집단과 다양한 지역에서 얻은 데이터를 사용하여 모델의 예측 능력을 향상시키는 방향으로 진행될 필요가 있으며 특히 각기 다른 생활 습관, 유전적 특성, 그리고 환경적 요인을 가진 인구집단에서 모델의 일반화 능력을 검증하는 데에 중요하게 작용할 것으로 판단된다. 최적의 성능을 발휘하는 모델을 개발하는 것 외에도, 모델의 예측과 결정 과정을 투명하고 이해하기 쉽게 만드는 것이 필수적이다. 의료분야에서의 결정은 생명과 직결되어 있기 때문에, 모델이 어떠한 기반 위에서 예측을 수행하는지를 명확히 파악하고, 그 과정을 검증할 수 있어야 한다. 이해 가능한 AI를 구현하는 다양한 기법들, 예를 들어 SHAP, LIME과 같은 모델 해석 기법이 연구에 적극 통합되어야 한다. 추후 확장된 당뇨 예측 모델 연구에서는 모델 개발 과정에서의 편향성을 최소화하고, 데이터 보안 및 프라이버시를 보장하는 기술적인 방안을 탐구하며, 윤리적인 기준과 법적인 규제를 반영한 실용적 모델 개발이 필요하다.

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지역지능화혁신인재양성(Grand ICT연구센터) 사업의 연구결과로 수행되었음" (IITP-2023-2020-0-01489)

References

[1] World Health Organization Diabetes: Keys Facts. https://www.who.int/news-room/fact-sheets/detail/diabetes, 2022

[2] Moustafa, Z., Evolutions de l'Intelligence Artificielle: Quels enjeux pour l'activite humaine et la relation Humain-Machine au travail? Activites , pp. 1-39, 2020

[3] Machine Learning. Java T Point; https://www.javatpoint.com

[4] Pankajray, Convolutional Neural Network (CNN) and Its Application－All You Need to Know, 2021.

[5] Islam, I.A. and Milon, M.I., Diabetes Prediction: A Deep Learning Approach, *International Journal of Information Engineering and Electronic Business* , 11, pp.21-27, 2019

[6] Zhou, H., Myrzashova, R. and Zheng, R. Diabetes Prediction Model Based on an Enhanced Deep Neural Network. EURASIP Journal on Wireless Communications and Networking, Article No. 148, 2020

[7] Naz, H. and Ahuja, S, Deep Learning Approach for Diabetes Prediction Using PIMA Indian Dataset. Journal of Diabetes & Metabolic Disorders , 19, pp.391-403, 2020

[8] Swapna, G., Soman, K.P. and Vinayakumar, R, Automated Detection of Diabetes Using CNN and CNN-LSTM Network and heart Rate Signals. Procedia, 2018

[9] Chowdary, P.B.K. and Kumar, R.U, An Effective Approach for Detecting Diabetes Using Deep Learning Techniques Based on Convolutional LSTM Networks. *International Journal of Advanced Computer Science and Applications*, 12, pp.519-525, 2021.

[10] Mat Jizat, J.A., Abdul Majeed, A.P.P., Ahmad, A.F., Taha, Z. and Yuen, E, *Evaluation of the Machine Learning Classifier in Wafer Defects Classification. ICT Express*, 7, pp.535-539. 2021.

[11] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., Johannes, R.S.: Using the ADAP

learning algorithm to forecast the onset of diabetes mellitus. In: *Annual Symposium on Computer Applications in Medical Care,* pp.261 - 265, 1998.

저자 소개

심현(Hyun Sim)

2002년 순천대학교 컴퓨터과학과 졸업(이학석사)
2009년 순천대학교 대학원 컴퓨터과학과 졸업(이학박사)
2020년~현재 순천대학교 스마트농업전공
2021년~현재 순천대학교 정보전산원 원장
2021년~현재 순천대학교 산학협력교육센터 센터장
2021년~현재 디지털트윈스마트시티연구소 소장
※ 관심분야 : 디지털트윈, 인공지능, 학습콘텐츠

김현욱(Hyun-Wook Kim)

2013년 건국대학교 경제학과 졸업 (경제학사)
2008년~2010년 HSBC FX TF
2010년~2014년 Intel Korea Image Division
2022년~현재 Kornerstone 대표
2023년~현재 순천대학교 미래산업인재양성사업단 위원, 디지털트윈연구소 연구원
※ 관심분야 : 인공지능, 머신러닝, 스마트공장