Smart Media Journal / Vol.14, No.12 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2025.14.12.146

# 3가지 유형의 XAI를 적용한 당뇨 예측 및 설명 시스템

## (A Diabetes Prediction and Explanation System using Three Types of XAI)

**안윤애*, 조한진****

(Yoon Ae Ahn, Han Jin Cho)

**요 약**

블랙박스 형태의 머신러닝 기반 당뇨 예측 시스템은 사용자에게 결과의 이유나 근거를 제시하지 못해 이해도와 신뢰도가 낮아지는 한계를 지닌다. 이를 개선하기 위해 본 논문은 3가지 유형의 설명 가능한 인공지능(XAI)을 적용한 당뇨 예측 및 설명 시스템을 설계·구현하였다. 피마 인디언 당뇨 데이터셋(PIDD)을 이용하여 중앙값 대체 방식으로 결측치를 처리하고, 9가지 분류 모델을 학습시켜 정확도, 정밀도, 재현율, F1-점수, ROC-AUC로 성능을 평가하였다. 최적 예측 모델에 LIME, SHAP, DiCE를 적용하여 각기 다른 설명 결과를 웹 기반 대시보드 형태로 시각화하였다. LIME과 SHAP은 예측 근거를 수치와 시각적으로 제시해 사용자의 이해를 높이며, DiCE의 반사실적 설명은 대안 시나리오를 통해 결과 해석을 돕는다. 세 가지 XAI를 하나의 예측 모델에 통합함으로써 당뇨 예측 시스템의 설명성과 신뢰성을 향상시켰으며, 본 연구는 향후 다른 의료 진단 시스템에도 XAI를 적용하는 기초 연구로 활용될 수 있을 것이다.

■ 중심어 : 설명가능한 AI ; 당뇨 예측 시스템 ; SHAP ; LIME ; 반사실적 설명

**Abstract**

In black-box machine learning-based diabetes prediction systems, users are not provided with reasons or explanations for the results, leading to low understanding and trust. To address this limitation, this study designs and implements a diabetes prediction and explanation system using three types of Explainable Artificial Intelligence (XAI) techniques. The Pima Indian Diabetes Dataset (PIDD) was used, and missing values were handled through median imputation. Nine classification models were trained and evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. The best-performing model was integrated with LIME, SHAP, and DiCE to visualize different types of explanations through a web-based dashboard. LIME and SHAP provide quantitative and visual representations of prediction rationale, enhancing user comprehension, while DiCE offers counterfactual explanations that present alternative scenarios for easier interpretation. By combining three XAI methods into one prediction model, the proposed system improves explainability and reliability in diabetes diagnosis. Furthermore, this study serves as a foundational reference for applying XAI techniques to other medical diagnostic systems.

■ keywords : Explainable AI ; Diabetes Prediction System ; SHAP ; LIME ; Counterfactual Explanation

## Ⅰ. 서 론

XAI(Explainable AI)는 일반적인 AI 시스템의 예측 결과에 대해 다양한 방법으로 그 이유를 수치값이나 그래프 형태로 제공한다. 이를 통해서

* 정회원, 한국교통대학교 컴퓨터공학과
** 정회원, 극동대학교 소프트웨어학과
접수일자 : 2025년 09월 20일
게재확정일 : 2025년 10월 17일
교신저자 : 조한진 e-mail : hanjincho@kdu.ac.kr

사용자는 시스템의 예측 결과에 대한 궁금증을 해소할 수 있으며, 최근에는 다양한 분야에서 XAI를 적용하고 있다. 특히 당뇨 예측과 같은 의료 진단 분야에서도 중요한 역할을 한다[1].

대표적인 XAI 기술로 LIME, SHAP, DiCE의 개념을 살펴본다. LIME(Local Interpretable Model-agnostic Explanations)은 시스템의 예측에 활용되는 사용자의 개별 데이터를 이용하여 선형 모델을 통해 국소적, 지역적(Local)으로 예측 결과에 대한 설명 값을 제공한다[2,3]. 지역적, 전역적(Global) 해석이 모두 가능한 기법에는 SHAP(SHapley Additive exPlanations)이 있다. 이 기법은 게임 이론을 토대로 하며 예측에 사용되는 특성, 피처(Feature)가 결과에 영향을 미치는 설명 값을 제공한다. 이때 개별 사용자의 피처를 사용하면 지역적인 해석이 가능하고, 모든 학습 데이터의 피처를 사용하면 전역적인 설명이 가능하다[2,3]. LIME과 SHAP의 단점은 예측 결과의 반대 경우에 대한 설명이 없다는 것이다. 반대 경우에 대한 설명 값을 제공하는 XAI를 반사실적 설명(Counterfactual Explanation)이라고 하며 대표적 기술이 DiCE이다. DiCE(Diverse Counterfactual Explanations)는 시스템의 예측 결과에 대한 반대 경우의 설명, 반사실적 시나리오를 제시한다. 이를 통해 의료 진단 분야에서는 특정 질병의 예측이 양성으로 나온 환자의 치료 방향에 대한 정보를 제공할 수 있다[4,5].

XAI는 당뇨 예측 시스템뿐만 아니라 다양한 의료 응용 시스템에 적용되고 있다. [6]에서는 유방암 연구에 LIME, SHAP을 적용하였고, [7]에서는 알츠하이머를 조기에 진단하는 연구에 XAI를 적용하였고, [8]에서는 파킨슨병을 조기에 진단하는 연구에 XAI를 적용하였다. 이 연구들에서는 사용자의 이해를 돕기 위해 그래프 등을 이용한 시각화를 활용하였다. 그러나 XAI 기법들은 설명 및 해석에 사용되는 이론적인 방법이 다르고, 결과 제공의 시각화도 서로 다르다. 이에 따라 한가지 기법만을 적용한 의료 응용 시스템의 예측 결과에 대해 완전한 사용자의 이해, 만족감을 주기는 어려운 상태이다[9,10].

따라서 이 논문에서는 3가지 유형의 XAI 기법을 적용한 웹 기반의 당뇨 예측 및 설명 시스템을 제안한다. 9가지의 분류 모델을 비교 분석하여 우수한 모델을 구축하고, 이 모델에 LIME, SHAP, DiCE 설명모델을 모두 적용하여 결과를 도출한다. 결과에 대한 사용자의 이해를 돕기 위해 웹 대시보드 형태로 시각화한 결과도 제공한다. 이 시스템은 하나의 예측 결과에 대해 지역적 설명, 전역적 설명, 반사실적 설명을 모두 제공함으로써 사용자의 신뢰도 및 이해도를 높일 것이라 기대한다. [9,10]의 연구에서도 2가지의 XAI를 사용한 시스템의 경우 사용자의 신뢰도가 다소 향상되었음을 제시하였다. 본 연구는 통합적 접근이 의료 예측 시스템의 사용자 이해를 돕고, 환자 개인에 대한 맞춤형 진료가 가능하도록 도와주는 밑바탕이 될 것이라 기대한다.

논문의 2장에서는 관련 연구를 기술하고 한계점과 해결 방안을 제시한다. 3장에서는 제안하는 당뇨 예측 시스템의 학습모델을 구축한다. 4장에서는 XAI를 적용한 당뇨 예측 및 설명 시스템을 설계하고 각 구성 모듈의 처리 과정을 설명한다. 제안 시스템의 실행 과정의 예시를 통해 3가지 유형의 XAI의 해석 결과 및 장단점을 분석하고, 5장에서 결론을 맺는다.

## Ⅱ. 관련 연구

### 1. 일반적인 AI 기반 당뇨 예측 시스템

블랙박스 형태의 일반적인 머신러닝 기법만을 적용한 당뇨병 예측 모델 연구 중, [11]에서는 피마 인디언 당뇨병 데이터 세트(PIDD)를 활용하여 5배 및 10배 교차 검증 방식으로 속성을 학습하여 딥 신경망을 사용한 당뇨병 진단 전략을 제안하였다. [12]에서는 PIDD와 규칙 기반 분류기 기술을 당뇨병 진단에 활용하기 위해 주성분 분석(PCA)을 추가로 사용하는 최소 규칙에 기반한

분류기를 제안하였다. [13]에서는 서포트 벡터 머신(SVM), 결정 트리(DT), 상관 분석을 사용하여 70%의 정확도로 PIDD를 예측하는 세 가지 중요한 요인을 제시하였다. [14]에서는 PIDD를 예측하기 위해 인공 신경망(ANN), 나이브 베이즈(NB), 결정 트리 및 딥 러닝(DL)의 4가지 머신러닝 알고리즘을 사용하여 당뇨병을 예측하는 방법론을 제시하였다. [15]에서는 PIDD에서 당뇨병 감지를 위한 세 가지 인기 있는 모델인 랜덤 포레스트(RF), 장단기 메모리(LSTM), 합성곱 신경망(CNN)에 대한 비교 연구를 수행하였으며, 그 결과로 LSTM 모델이 85%의 최고 정확도를 달성하여 당뇨병을 예측하는 정확한 방법으로 효과가 있음을 보였다. [16]에서는 K-최근접 이웃(KNN) 알고리즘을 사용하여 PIDD의 당뇨병 예측을 연구하였다. 실험과 평가를 통해 임상 변수에 따라 당뇨병 발병을 정확하게 예측하는 데 KNN 기법이 성능이 좋음을 제시하였다. 지금까지 검토한 당뇨 예측 연구들은 블랙박스 모델의 AI 기법들을 사용하여 예측 결과에 대한 설명성을 전혀 제시하지 못하는 한계가 있다.

## 2. XAI를 적용한 당뇨 예측 시스템

기존 블랙박스 모델의 한계점을 해결하기 위해 당뇨 예측 모델에 XAI를 적용한 연구 중, [17]에서는 방글라데시의 여성 환자에 대한 개인 데이터 세트를 ADASYN, XGBoost 분류기를 사용한 자동 당뇨병 예측 시스템을 개발하였으며, LIME과 SHAP을 구현하여 모델이 최종 결과를 예측하는 방식을 제시하였다. [18]에서는 반사실적 설명을 기반으로 당뇨병의 1년 위험을 줄이기 위한 개인화된 권장 사항을 생성하기 위해 5,582명의 의료 기록에서 추출한 10개의 바이오마커를 분석, 지원 벡터 머신(SVM) 분류기를 사용하여 당뇨병을 예측하였다. 여기에 XAI의 반사실적 설명을 사용하여 당뇨병 위험이 높은 환자에 대해 몸무게, 혈당, 혈압 등의 수치를 낮추도록 권장하고 있다. [19]에서는 PIDD를 활용하여 랜덤 포레스트, 의사결정 트리, 그래디언트 부스팅, 신경망 등의 머신러닝 알고리즘을 기반으로 당뇨병 발병을 예측하였고, 예측 모델에서 해석 가능성을 SHAP을 통해서 설명하고 있다. [20]에서는 고성능의 해석 가능한 당뇨병 예측 모델을 찾기 위해 KNN, 10배 교차 검증, 익스트림 그래디언트 부스팅(XGBoost) 알고리즘을 활용하여 당뇨병 예측 모델을 제시하고, 결과의 이해도를 높이기 위해 SHAP, LIME을 사용한 전역 및 지역적인 설명 정보를 제공하였다. [21]에서는 로지스틱 회귀 아키텍처 기반 머신러닝 모델에 대한 설명을 제시하기 위해, 미국 질병통제예방센터(CDC)에 제출된 당뇨병 환자의 253,680개 설문 응답 데이터 세트를 사용하여 LIME, SHAP을 통해 로지스틱 회귀(Logistic Regression) 및 랜덤 포레스트 기반 모델의 예측 결과에 대한 설명성을 추가하였다. [22]에서는 PIDD에 비전문가도 사용이 가능한 자동화된 머신러닝(AutoML)을 활용하여 학습모델을 구축한 후 XAI로 LIME과 SHAP을 적용하였다. 설명 결과는 Python의 Streamlit 앱을 통해 시각화하여 제시하였다. [23]에서는 미국 질병통제예방센터의 BRFSS 데이터에 대화형 AI 학습모델을 구축하여, 모델의 예측 결과에 LIME과 SHAP의 설명모델을 활용하여 설명 결과를 웹 대시보드로 제시하였다.

## 3. 기존 연구의 한계점 및 해결 방안

XAI 기반의 당뇨 예측 모델은 SHAP과 LIME을 적용한 사례가 대부분이다. 그러나 최근에는 이 2가지 기법의 부족한 점을 보완하기 위해서 반사실적 설명 기법이 다양하게 활용되고 있다. 특히 DiCE를 이용한 반사실적 설명 기법은 당뇨 예측 모델에서 “어떤 피처의 값을 변화시키면 예측 결과를 바꿀 수 있는가?”를 설명하여 사용자에게 당뇨를 예방하기 위한 행동 목표를 제시할 수 있다[4,5]. 실제로 2023년도 연구[24]에서는 임상 전문가를 대상으로 DiCE를 적용한 대시보드 프로토타입을 평가하여, 반사실적 설명의 제공

값이 실질적으로 일반 사용자에게 받아들여질 수 있음을 제시하였다.

지금까지 검토한 XAI 기반의 당뇨 예측에 관한 선행 연구들에서는 SHAP 또는 LIME 기법 중 1가지 또는 2가지를 적용한 사례가 대부분이며, DiCE를 포함한 3가지 이상의 적용 사례 및 웹 애플리케이션 대시보드 형태의 프로토타입은 거의 없는 실정이다.

따라서, 본 논문에서는 기존 당뇨 예측 모델 연구의 설명성을 강화하기 위해 3가지 XAI 기술인 LIME, SHAP, DiCE를 동일한 당뇨 예측 모델에 모두 적용하여, 예측 결과에 대한 3가지의 설명을 제시한다. 아울러 웹 기반 대시보드를 통해 로컬(지역적 설명), 글로벌(전역적 설명), 반사실적 설명(시나리오)의 해석 결과를 제공하여, 사용자가 다양한 형태로 당뇨 예측 시스템의 결과를 쉽게 이해할 수 있는 시스템을 구현한다.

## Ⅲ. 당뇨 예측 시스템의 학습 모델 구축

이 논문에서는 당뇨 예측 모델 개발을 위해 Pima Indians Diabetes Dataset(PIDD)을 활용하였다. PIDD는 미국 애리조나주에 거주하는 피마 인디언 여성 768명의 건강 데이터를 포함한 공개 벤치마크 데이터 세트로, 혈당, BMI, 나이 등 8개의 피처와 당뇨 여부를 나타내는 Outcome 변수로 구성된다. 이러한 구성은 많은 선행 연구에서 당뇨 예측에 이미 사용되었다[25].

8개의 기본 피처에는 'Pregnancies' 임신 횟수, 'Glucose' 혈당, 'BloodPressure' 이완기 혈압, 'SkinThickness' 삼두근 피부 두께, 'Insulin' 혈청 인슐린 수치, 'DiabetesPedigreeFunction' 당뇨병 가족력 함수, 'BMI' 체질량지수, 'Age' 나이가 있다. 1개의 Outcome 변수는 당뇨병 여부로 1(당뇨) 또는 0(비 당뇨)을 나타낸다[26].

그림 1은 PIDD 데이터를 활용한 당뇨 예측 모델의 구축 과정을 나타내며, 다음과 같이 다섯 단계로 처리하였다.

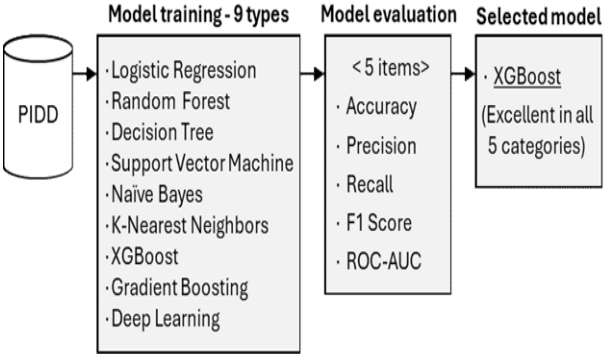


그림 1. 예측 모델의 선정 과정

① 데이터 전처리 : PIDD의 8가지 피처 중에서 'Glucose' 혈당수치, 'BloodPressure' 이완기 혈압, 'SkinThickness' 삼두근 피부 두께, 'Insulin' 인슐린 수치, 'BMI' 체질량지수 값에는 0 값이 포함되어 있으며, 이는 실제로 측정되지 않았거나 누락된 데이터이다. 이 0 값을 각 피처의 중앙값으로 대체하여 데이터의 완전성을 확보하였다. 데이터 스케일링 처리는 StandardScaler 함수의 fit_transform과 transform을 사용하였다.

② 후보 학습모델 비교 실험 : 실험을 위해 다음과 같이 9개의 분류 모델을 테스트하여 성능이 좋은 모델 하나를 선택하였다. 로지스틱 회귀, 랜덤 포레스트, 결정 트리, 서포트 벡터 머신, 나이브 베이즈, K-최근접 이웃, 익스트림 그래디언트 부스팅, 그래디언트 부스팅, 딥 러닝 모델이 사용되었다.

③ 하이퍼파라미터 값 설정 : 9개 학습모델의 성능을 위해 각각의 하이퍼파라미터를 찾아야 한다. 이를 위해 하이퍼파라미터가 없는 Naive Bayes를 제외한 8가지 모델에는 GridSearchCV 함수에 5배 교차 검증을 사용하였다.

④ 5가지 평가지표 활용 : 정확도(Accuracy)는 예측 모델의 모든 결과 중 원 데이터와 일치하는 결과의 비율을 나타낸다. 정밀도(Precision)는 시스템 예측 모델의 모든 예측 결과 중 양성으로 나온 값들이 실제로 양성인 비율을 계산한 것이다. 재현율(Recall)은 원 데이터에서 양성인 값들이 예측 모델의 결과에서도 양성인 비율을 계산한

것이다. F1-점수는 정밀도 값과 재현율 값을 조화평균으로 계산한 것으로 두 가지 값의 균형을 나타낸다. ROC-AUC는 ROC(수신자 조작 특성 곡선)와 AUC(곡선 아래 면적) 값을 사용하여 종합적으로 분류 모델을 평가하며, 특히 데이터가 불균형한 경우에 적합하다.

⑤ 최종 학습모델 선정 : 여러 모델의 성능과 해석 용이성을 비교하여 최종 모델을 결정하였다. 당뇨병 예측 모델의 경우 정확도, F1-점수, ROC-AUC의 평가지표를 종합적으로 고려해야 한다[15,17]. 이에 따라 선택된 최종 모델은 평가 항목에서 우수한 성능을 보인 XGboost이다. 이 모델은 강력한 예측 성능과 효율적인 계산 속도로 인해 널리 사용되는 알고리즘이다[17,23].

| | Model | Accuracy | Precision | Recall | F1 Score | ROC AUC |
|---|---|---|---|---|---|---|
| 0 | Logistic Regression | 0.772727 | 0.739130 | 0.596491 | 0.660194 | 0.864532 |
| 1 | Random Forest | 0.811688 | 0.780000 | 0.684211 | 0.728972 | 0.866341 |
| 2 | Decision Tree | 0.805195 | 0.721311 | 0.771930 | 0.745763 | 0.797251 |
| 3 | Support Vector Machine | 0.772727 | 0.789474 | 0.526316 | 0.631579 | 0.868692 |
| 4 | Naive Bayes | 0.746753 | 0.660714 | 0.649123 | 0.654867 | 0.824561 |
| 5 | K-Nearest Neighbors | 0.746753 | 0.695652 | 0.561404 | 0.621359 | 0.832339 |
| 6 | XGBoost | 0.850649 | 0.793103 | 0.807018 | 0.800000 | 0.870863 |
| 7 | Gradient Boosting | 0.779221 | 0.725490 | 0.649123 | 0.685185 | 0.832519 |
| 8 | Deep Learning | 0.792208 | 0.777778 | 0.614035 | 0.686275 | 0.860011 |

그림 2. 9가지 학습모델의 평가 결과

그림 2는 9가지 학습모델을 평가한 결과이다. XGBoost는 정확도(85.1%), 정밀도(79.3%), 재현율(80.7%), F1-점수(80.0%), ROC-AUC(87.1%)에서 타 모델보다 전반적으로 뛰어난 성능을 보인다. 특히 재현율, F1-점수, ROC-AUC 지표에서 모두 높은 수치를 기록하여, 양성과 음성 구분 성능이 고르게 우수하고 불균형 클래스 상황에서도 우수한 예측력을 보인다. 아울러 기존의 유사한 연구[25,27]에서 보여준 XGBoost의 성능 결과인 정확도 최대 85%, ROC-AUC 최대 91%와 유사한 수준임이 확인되었으며, PIDD 기반 예측 모델의 타당성을 뒷받침한다.

## Ⅳ. XAI 기반 당뇨 예측 및 설명 시스템

### 1. 시스템 구조

일반적인 머신러닝 기반의 당뇨 예측 모델에 3가지 유형의 XAI 설명모델을 추가하여, 웹 기반 클라이언트에서 실행이 가능하도록 설계한 시스템의 구조는 그림 3과 같다.

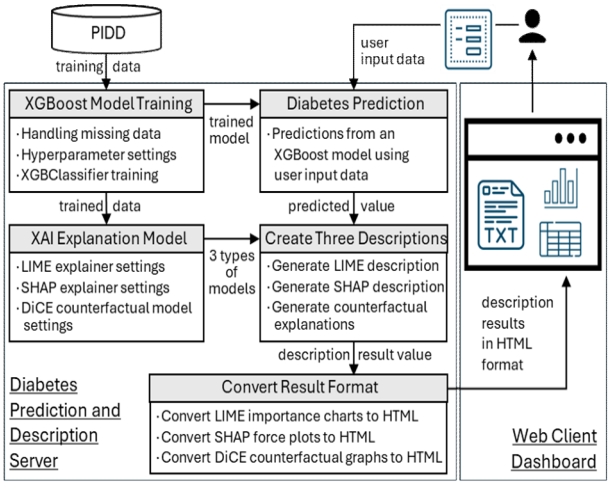


그림 3. 당뇨 예측 및 설명 시스템 구조

각 구성 요소의 처리 흐름은 다음과 같다.

- XGBoost Model Training 모듈: 앞의 3절에서 선정한 XGBoost 학습모델에 PIDD를 학습용 데이터로 사용하여 모델을 학습시킨다. 이를 위해 결측값을 처리하고 하이퍼파라미터를 조정하여 XGBClassifier 함수를 실행한 후 학습 완료된 XGBoost 모델을 도출한다.
- XAI Explanation Model 모듈: 3가지 유형의 설명모델을 구축한다. 학습 데이터를 사용하여 LIME은 LimeTabularExplainer 함수를 사용하고, SHAP은 shap.Explainer 함수를 사용하고, DiCE는 dice_ml.Model 함수를 사용하여 반사실적 설명모델을 생성한다. 이들 설명모델은 서로 다른 시점에서 각각 활용된다.
- Diabetes Prediction 모듈: 웹 클라이언트에서 사용자가 당뇨 예측을 위해 입력한 8가지의 특성값(Feature Value)을 받아서, XGBClassifier 함수로 생성한 예측 모델을 통해 당뇨 예측값(Predicted Value)을 생성한다. 예측 결과는 1(양

성:당뇨), 0(음성:비 당뇨)으로 도출된다.

- Create Three Descriptions 모듈: LIME은 설명모델에 explain_instance 함수를 사용하여 개별 인스턴스에 대한 지역적 설명 값을 도출한다. SHAP은 설명모델에 shap_explainer 함수를 사용하여 개별 인스턴스에 대한 지역적 설명 값과 모든 학습 데이터(X_train)에 대한 전역적 특성 중요도 값을 도출한다. DiCE는 반사실적 설명에 적용될 피처들을 선택하고 이들의 변경 값의 범위를 최소값과 최대값의 비율로 설정한다. 이후 dice.generate_counterfactuals 함수를 사용하여 모델 기반의 반사실적(Counterfactual) 시나리오 설명 값을 도출한다.
- Convert Result Format 모듈: LIME의 explain_instance 함수를 사용하여 생성한 설명 값과 dict 그래프 함수의 결과를 html 형태로 변환하여 Flask의 render_template 함수를 통해 클라이언트에게 결과를 전달한다. SHAP의 shap_explainer 함수로 생성한 설명 값과 shap의 force_plot, plots.force, plots.waterfall 그래프 함수로 작성된 결과를 html로 변환하여 결과를 전달한다. DiCE의 dice.generate_counterfactuals 함수로 도출된 반사실적 설명 값과 matplotlib의 pyplot 함수로 작성한 그래프 결과를 html 형태로 변환하여 결과를 전달한다.
- Web Client Dashboard 모듈: Flask 웹 서버의 render_template 함수를 통해 모든 결과를 result.html 파일로 전송받는다. 웹브라우저를 통해서 클라이언트(사용자)는 당뇨 예측 결과 및 LIME, SHAP, DiCE의 설명 값을 텍스트, 챠트, 반사실적 시나리오 예시 등이 포함된 결과 형태로 제공받는다.

## 2. 구현 시스템의 실행

제안 시스템은 Python 3.7과 Flask 웹 프레임워크로 구현하였다. 구현 시스템의 실행 과정은 PIDD의 테스트 데이터 중 실제 당뇨로 판명된 사용자의 데이터를 활용하여 설명한다.

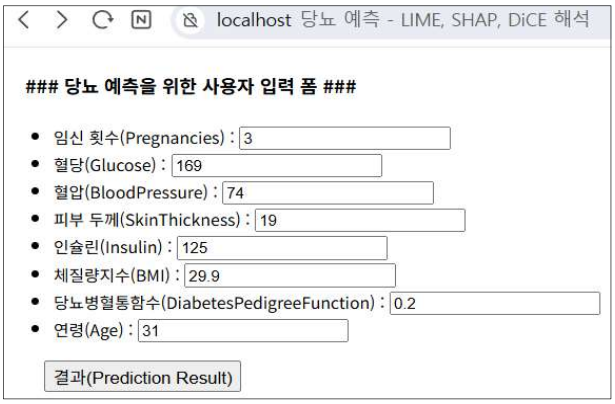


그림 4. 당뇨 예측을 위한 사용자 피처 입력

그림 4와 같이 사용자의 8개의 피처 값을 입력하고 결과 버튼을 클릭한다. 이후 사용자가 첫 번째로 리턴 받는 결과는 그림 5와 같이 XGBoost 학습모델이 예측한 당뇨 예측 결과 정보이다.

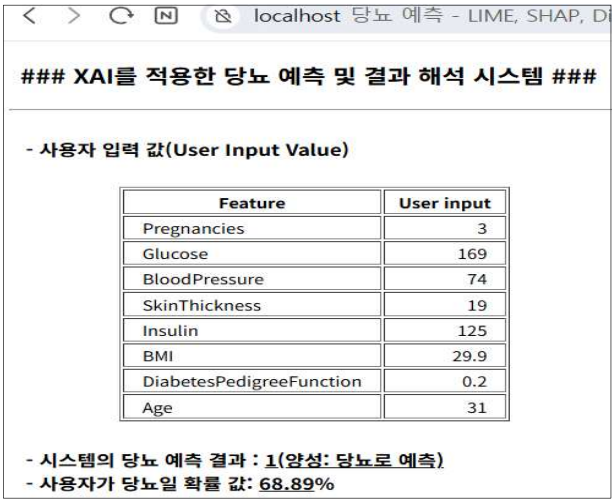


그림 5. XGBoost 모델의 당뇨 예측 결과 정보

그림 5는 XGBoost 학습모델이 예측한 당뇨 예측 결과 정보이다. 사용자가 입력한 피처 값이 표 형태로 정리되어 제공되고, 예측 결과와 확률도 함께 표시된다. 예제에서 시스템의 당뇨 예측 결과는 1(양성: 당뇨로 예측)이고, 사용자가 당뇨일 확률값은 68.89%이다. 입력 데이터, 예측 결과, 확률을 한 번에 보여주어 사용자의 이해를 돕는다. 당뇨 예측 결과에 대한 설명, 해석은 다음의 3가지 유형으로 순서대로 확인할 수 있다.

### 가. LIME을 적용한 지역적 설명

LIME은 개별 예측을 설명하는 기법이다. 사용자가 입력한 개별 피처 데이터를 활용하여 선형 모델을 생성한 후 설명 값을 제공한다. 이를 로컬(지역적) 해석이라고 한다.

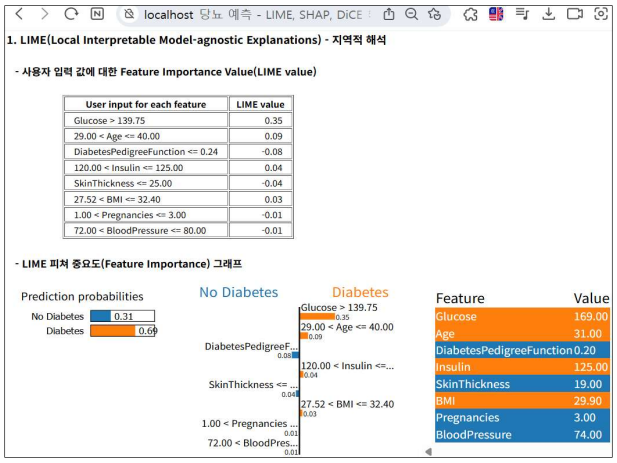


그림 6. LIME의 로컬(지역적) 설명 값과 챠트

그림 6은 사용자의 개별 입력 피처에 대한 분석을 LIME value를 통해서 표와 챠트 형태로 제공한 것이다. 표의 LIME value가 클수록 예측 결과(양성: 당뇨로 예측)에 더 많은 영향을 끼침을 의미한다. 챠트에서 피처별로 예측에 기여한 정도를 파악하는 방법은 Diabetes 아래의 막대가 오른쪽으로 길게 표시된 것을 확인하면 된다. 챠트에서 Glucose, Age, Insulin, BMI 순으로 예측 결과에 대한 기여도가 높음을 알 수 있다.

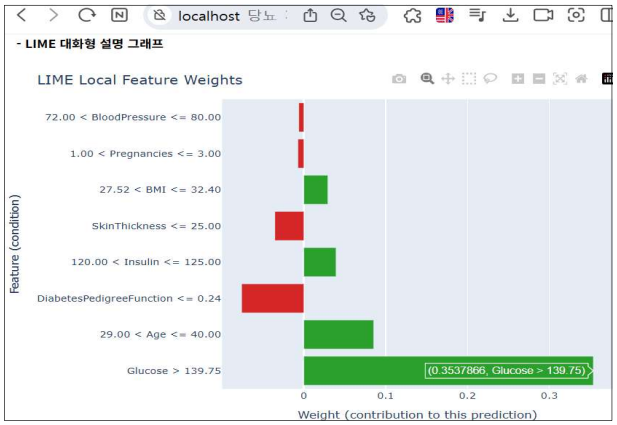


그림 7. LIME의 대화형 설명 그래프

그림 7은 LIME의 대화형 설명 그래프이다. 세로축의 각 피처들이 가로축의 0을 중심으로 오른쪽(양수 값)으로 길게 나타나면 예측 결과에 대한 영향력(Weight)이 크다고 해석하면 된다. 막대 위에 마우스를 올리면 해당 피처의 Weight 수치를 확인할 수 있다.

나. SHAP을 적용한 지역적/전역적 설명

SHAP은 사용자의 개별 데이터를 활용한 로컬(지역적) 설명, 모든 학습 데이터 및 모든 피처를 활용한 글로벌(전역적) 해석이 모두 가능한 기법이라서 LIME의 단점을 다소 보완할 수 있다.

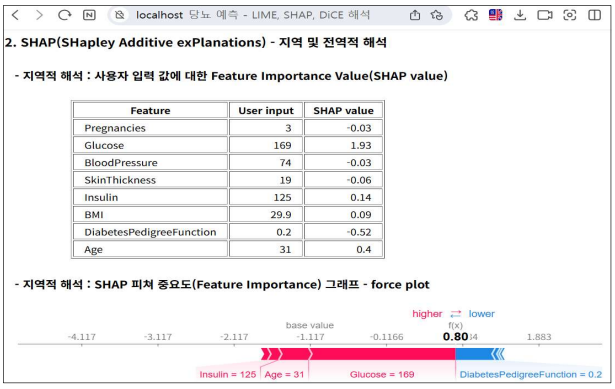


그림 8. SHAP의 로컬(지역적) 설명 값과 챠트

그림 8에서 표에 출력된 SHAP value가 큰 피처일수록 예측 결과에 영향력을 많이 미치는 것이다. Glucose, Age, Insulin, BMI 순으로 기여도가 높으며, LIME의 결과와도 일치한다. 표 아래의 force plot은 시각적인 이해를 돕기 위한 것으로 higher 글자와 동일한 색상의 막대 길이가 길수록 영향력이 높은 피처로 해석한다.

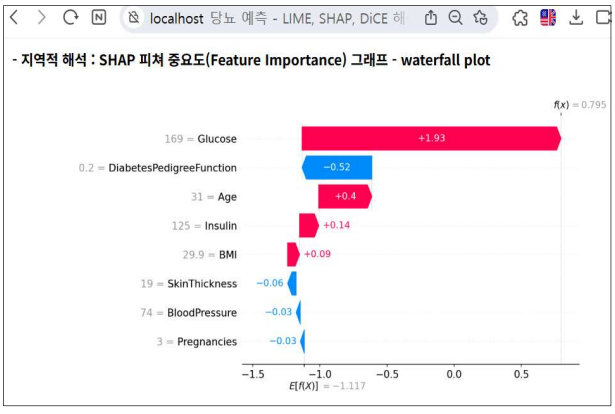


그림 9. SHAP의 로컬(지역적) 워터폴 그래프

그림 9는 SHAP의 지역적 피처 중요도를 waterfall plot으로 보여준다. 사용자는 그림 8의 force plot과 비교하면서 좀 더 직관적으로 예측 결과에 대한 이유를 이해할 수 있다.

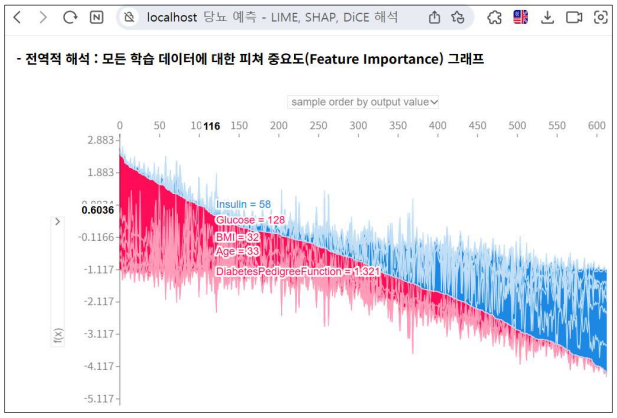


그림 10. 모든 학습 데이터에 대한 글로벌(전역적) 챠트

그림 10은 SHAP의 전역적인 해석을 보여주는 결과 그래프이다. PIDD를 활용한 전체 학습 데이터들의 피처 중요도 변화를 한 번에 보여주고 있다. 챠트의 아래쪽 영역(빨간색)은 Outcome 피처가 양성(1)인 학습 데이터를 나타내고, 위쪽 영역(파란색)은 음성(0)인 학습 데이터를 나타낸다. 마우스를 이동하면서 각 개별 학습 데이터의 피처 값과 예측 결과를 한눈에 이해할 수 있다.

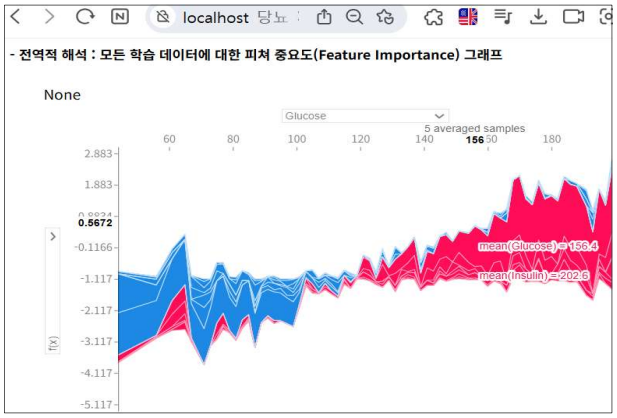


그림 11. 모든 학습 데이터에 대한 Glucose 챠트

그림 11은 그림 10의 옵션(콤보박스)에서 Glucose 피처를 선택하여 변경된 챠트이다. 모든 학습 데이터의 Glucose 값을 통해 영향력을 확인할 수 있다. 가로축에서 오른쪽(빨간색)으로 갈수록 Glucose의 영향력이 크고, Outcome 피처가 양성(1)인 학습 데이터를 나타내는 것이다.

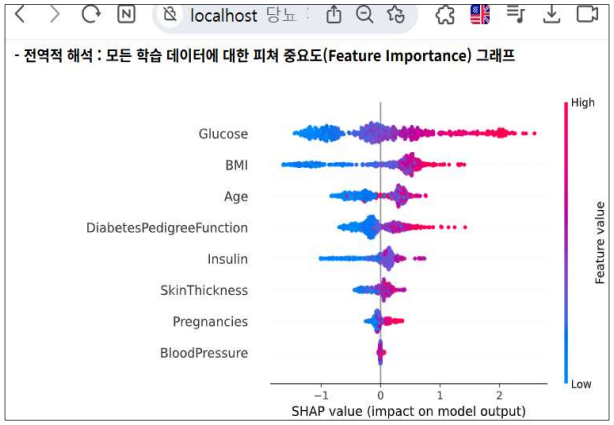


그림 12. 모든 학습 데이터에 대한 글로벌(전역적) 챠트 (8개의 피처 중심으로 중요도 확인)

그림 12는 모든 피처에 대한 중요도의 순위를 8개 피처 각각에 대한 SHAP summary plot 형태로 보여준다. Outcome이 1(양성)인 모든 학습 데이터를 대상으로 피처 중요도가 가장 높은 것은 Glucose, BMI, Age의 순으로 나타나고 있다.

### 다. DiCE를 적용한 반사실적 설명

DiCE는 학습모델의 예측 결과를 반대로 만들기 위한 피처 값들을 제공하며, 이를 반사실적 설명이라고 한다. 만약, 어떤 사용자의 예측 결과가 1(양성)인 경우, 반사실적 설명은 0(음성)이 될 수 있는 피처 값들을 제공한다.

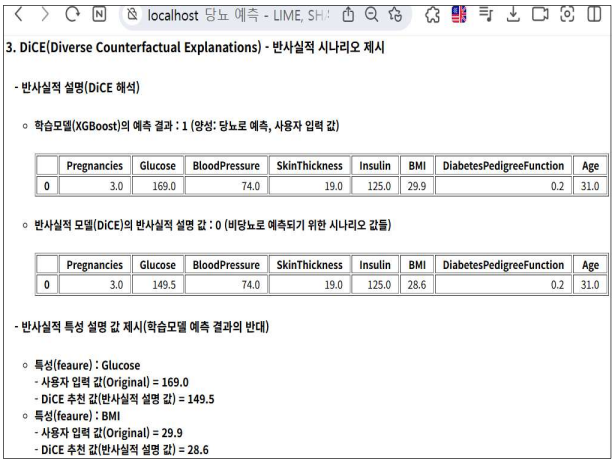


그림 13. DiCE의 반사실적 설명(시나리오) 결과

그림 13에서 DiCE는 what-if 설명을 제공한다. 현재 사용자의 예측 결과가 1(양성)이므로, 반대 0(음성)으로 만들기 위해서는 Glucose 피처는 149.5, BMI 피처는 28.6 정도가 필요하다고 제시하고 있다. 그림 14는 DiCE의 반사실적 설명 값의 비교 챠트로서 시각적인 설명을 지원한다.

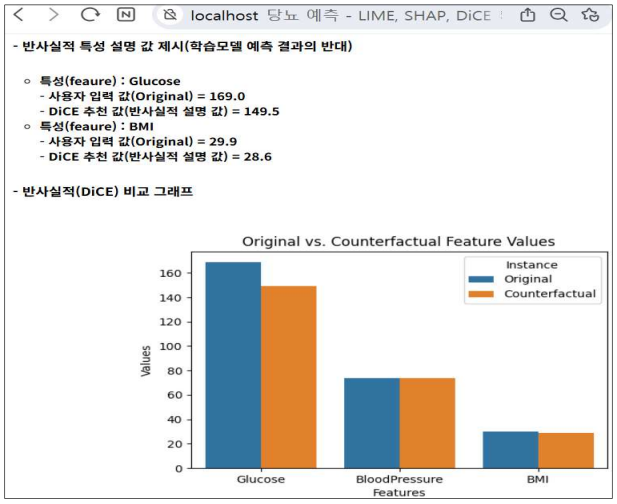


그림 14. DiCE의 반사실적 설명 값의 비교 챠트

### 3. 결과 분석 및 고찰

구현 시스템의 실행 결과를 토대로 3가지 유형의 설명 기법에 대해 간략하게 분석해 본다.

#### 가. 3가지 XAI의 특성 분석

LIME은 매우 간단한 선형 모델을 설명에 적용하는 방법이다. 이는 개별 데이터에 대한 로컬(지역적) 해석 결과만을 제공하므로, 글로벌(전역적) 해석 결과가 부족한 단점이 있다. 따라서 SHAP의 설명을 추가하여 보완하였다. SHAP은 전역적 해석 결과를 여러 가지 형태의 그래프로 제공하여 설명 결과의 다양성을 높일 수 있었다. 그러나 SHAP은 당뇨로 예측된 사용자가 비당뇨로 예측되기 위한 반사실적(반대) 시나리오와 같은 결과를 제시하지는 못하는 단점이 있다. 이를 보완하기 위해서 DiCE를 추가로 사용하였다. DiCE는 모델의 예측 결과를 반대로 변경시키는 데 필요한 값을 제시하는 기법이다. what-if 시나리오 즉, 사용자가 원하는 결과를 얻기 위해 입력 피처들을 어떤 값으로 변경해야 하는지 몇 가지 대안을 제시한다. 그렇지만 DiCE의 결과로 제시된 반대(대안) 값들이 현실성이 있는지, 사용자가 실행할 수 있는 대안인지에 대한 추가적인 검토가 필요하다. 이 점은 향후 연구에서 구체적으로 진행할 필요가 있다.

#### 나. 연구의 한계점 및 향후 연구

이 논문에서는 공개적으로 사용이 가능한 데이터 세트인 PIDD를 사용하였다. 이는 피마 인디언 여성에 한정된 소규모 데이터이므로 일반인으로의 확장성에는 다소 애로점이 있다. 아울러 시스템의 설명성에 대한 사용자의 만족도 평가를 수행하지 못한 한계점이 있다. 따라서 향후에는 국내에서 수집된 공개 사용이 가능한 데이터 세트를 확보하여 테스트할 필요가 있으며, 사용자 만족도 평가를 위한 방법을 연구할 계획이다.

## V. 결 론

이 논문에서는 기존의 블랙박스 기반의 당뇨 예측 시스템의 한계점을 극복하기 위해 3가지 유형의 XAI 기법을 적용한 시스템을 설계하고 웹 대시보드로 구현하였다. 일반적으로 당뇨 예측 모델에 적용되는 XAI 기법의 종류에 따라 결과에 대한 해석 방법이 달라질 수 있다. 이를 보완하기 위해서 하나의 예측 결과에 대해 서로 다른 3가지의 해석 기법 LIME, SHAP, DiCE 모델을 동시에 적용하는 방법을 선택하였다. 이러한 다양한 XAI 기법의 적용을 통해 당뇨병 예측 모델의 예측 결과를 다각도로 해석할 수 있었으며, 각 기법의 특성과 장단점을 파악하였다. 이를 통해 의료 분야에서 환자의 질병 진단에 사용되는 머신러닝 예측 모델의 이해도, 신뢰성, 투명성을 높이는 데 기여할 수 있을 것이며, 의료 XAI 응용의 실용적 레퍼런스로 활용이 가능할 것이다.

그러나, 아직 이 연구에서는 제안 시스템의 결과에 대한 사용자의 만족도를 측정하는 방법을

적용하지 못하는 한계점이 있다. 따라서 향후 연구에서는 웹 기반 사용자의 만족도를 평가하는 방법을 구체적으로 연구할 필요성이 있다.

## REFERENCES


[1] F. Lecue, "Industrial applications of XAI: Tutorial on Explainable AI," *Proceedings of The Thirty-Third AAAI Conference on Artificial Intelligence, Tutorial Slides*, Jan., 2019.

[2] A.B. Arrieta, N. Diaz-Rodriguez, J.D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82-115, 2020.

[3] M. Mersha, K. Lam, J. Wood, A. AlShami, and J. Kalita, "Explainable Artificial Intelligence: A Survey of Needs, Techniques, Applications, and Future Direction," *Neurocomputing*, vol. 599, article no. 128111, pp. 1-31, Sep. 2024.

[4] R.K. Mothilal, D. Mahajan, C. Tan, and A. Sharma, "Towards Unifying Feature Attribution and Counterfactual Explanations: Different Means to the Same End," *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society, AIES2021*, pp. 306-317, 2021.

[5] S. Singla, M. Eslami, B. Pollack, S. Wallace, and K. Batmanghelich, "Explaining the black-box smoothly- A counterfactual approach," *Medical Image Analysis*, vol. 84, article no. 102721, pp. 1-13, 2023.

[6] A. Sathyan, A.I. Weinberg, and K. Cohen, "Interpretable AI for bio-medical applications," *Complex Engineering Systems*, vol. 2, no. 4, article no. 18, pp. 1-18, Dec. 2022.

[7] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: A systematic review on the application of LIME and SHAP in Alzheimer's disease detection," *Brain Informatics*, vol. 11, article no. 10, pp. 1-29, Apr. 2024.

[8] P.R. Magesh, R.D. Myloth, and R.J. Tom, "An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery," *Computers in Biology and Medicine*, vol. 126, article no. 104041, pp. 1-21, Nov. 2020.

[9] R. Rosenbacke, A. Melhus, M. McKee, and D. Stuckler, "How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review," *Journal of Medical Internet Research AI*, vol. 3, article no. e53207, pp. 1-10, Oct. 2024.

[10] L. Pantanowitz, M. Hanna, J. Pantanowitz, J. Lennerz, W.H. Henricks, P. Shen, B. Quinn, S. Bennet, and H.H. Rashidi, "Regulatory Aspects of Artificial Intelligence and Machine Learning," *Modern Pathology*, vol. 37, no. 12, article no. 100609, pp. 1-8, Dec. 2024.

[11] S.I. Ayon and M.M. Islam, "Diabetes Prediction: A Deep Learning Approach," *International Journal of Information Engineering and Electronic Business (IJIEEB)*, vol. 11, no. 2, pp. 21-27, Mar. 2019.

[12] A.T. Reddy and M. Nagendra, "Minimal Rule-Based Classifiers using PCA on Pima-Indians-Diabetes-Dataset," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 12, pp. 4414-4420, Oct. 2019.

[13] S. You and M.S. Kang, "A Study on Methods to Prevent Pima Indians Diabetes using SVM," *Korean Journal of Artificial Intelligence*, vol. 8, no. 2, pp. 7-10, 2020.

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

[15] A. Mousa, W. Mustafa, R.B. Marqas, and S.H.M. Mohammed, "A Comparative Study of Diabetes Detection Using The Pima Indian Diabetes Database," *Journal of the University of Duhok*, vol. 26, no. 2, pp. 277-288, 2023.

[16] S.R. Mishra and S. Dash, "Predictive Analysis On Diabetes Detection Using Pima Indian Diabetes Dataset," *International Journal Research and Analytical Reviews*, vol. 11, no. 2, pp. 587-599, 2024.

[17] I. Tasin, T.U. Nabil, S. Islam, and R. Khan, "Diabetes Prediction using Machine Learning and Explainable AI Techniques," *Healthcare Technology Letters*, vol. 10, no. 1-2, pp. 1-10, Dec. 2022.

[18] M. Lenatti, A. Carlevaro, A. Guergachi, K. Keshavjee, M. Mongelli, and A. Paglialonga, "A novel method to derive personalized minimum viable recommendations for type 2 diabetes prevention based on counterfactual explanations," *Public Library of Science ONE*, vol. 17, no. 11, article no. e0272825, pp. 1-24, Nov. 2022.

[19] R. Alam and M. Atif, "Unveiling Diabetes

Predictions: Bridging Complexity and Clarity through Interpretable AI and SHAP Insights," *Proceedings of 4th International Conference on Data Analytics for Business and Industry (ICDABI)*, pp. 258-261, Oct. 2023.
[20] Y. Zhao, J.K. Chaw, M.C. Ang, M.M. Daud, and L. Liu, "A Diabetes Prediction Model with Visualized Explainable Artificial Intelligence (XAI) Technology," *Advances in Visual Informatics, Lecture Notes in Computer Science*, vol. 14322, pp. 648-661, 2023.
[21] S. Ahmed, M.S. Kaiser, M.S. Hossain, and K. Andersson, "A Comparative Analysis of LIME and SHAP Interpreters with Explainable ML-Based Diabetes Predictions," *IEEE Access*, vol. 13, pp. 37370-37388, 2024.
[22] R. Hasan, V. Dattana, S. Mahmood, and S. Hussain, "Towards Transparent Diabetes Prediction: Combining AutoML and Explainable AI for Improved Clinical Insights," *Information*, vol. 16, article no. 7, pp. 1-31, 2025.
[23] U. Allani, "Interactive Diabetes Risk Prediction Using Explainable Machine Learning: A Dash-Based Approach with SHAP, LIME, and Comorbidity Insights," arXiv preprint arXiv:2505.05683, May 2025.
[24] M.H. Lee and C.J. Chew, "Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW2, article no. 369, pp. 1-22, Oct. 2023.
[25] Z. Zhang, "Comparison of Machine Learning Models for Predicting Type 2 Diabetes Risk Using the Pima Indians Diabetes Dataset," *Journal of Innovations in Medical Research*, vol. 4, no. 1, pp. 65-71, Feb. 2025.
[26] Kaggle, "Pima Indians Diabetes Database," https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database(accessed Jan., 14, 2025).
[27] U.H. Okwudili, O.O. Ukachukwu, V.C. Chijindu, M.O. Ezea, and B. Ishaq, "An improved performance model for artificial intelligence-based diabetes prediction," *Journal of Electrical Systems and Information Technology*, vol. 12, article no. 25, pp. 1-30, 2025.

## 저 자 소 개

**안윤애(정회원)**

1996년 충북대학교 전자계산학과 석사 졸업.
2003년 충북대학교 전자계산학과 박사 졸업.
2003년~현재 한국교통대학교 컴퓨터공학과 교수
<주관심분야 : 헬스케어시스템, 인공지능, SW응용>

**조한진(정회원)**

1999년 한남대학교 컴퓨터공학과 석사 졸업
2002년 한남대학교 컴퓨터공학과 박사 졸업
2002년~현재 극동대학교 소프트웨어학과 교수
<주관심분야 : 인공지능, 정보보호, 클라우드, 빅데이터>