International Journal on Advanced Science Engineering Information Technology

Vol.15 (2025) No. 2
ISSN: 2088-5334

Diabetes Early Prediction Using Machine Learning and Ensemble Methods

Hyung-Ho Ha [a], Hangun Kim [a], Young Hyun Yu [a], Hyun Sim [b,*]

[a] *Department Pharmacy, Sunchon National University, Republic of Korea*
[b] *Department Smart Agriculture, Sunchon National University, Republic of Korea*
*Corresponding author: *simhyun@scnu.ac.kr*

Abstract—**This study aims to develop and validate an enhanced early prediction model for diabetes utilizing machine learning and ensemble techniques, aimed at addressing the rapid increase in diabetes prevalence and the associated healthcare burden. Leveraging diverse datasets, including the Pima Indian Diabetes Dataset, electronic health records from local hospitals, and wearable device data, this research employs a variety of innovative methods. Generative Adversarial Networks (GAN) are used for data augmentation to address class imbalances, while SHAP (Shapley Additive exPlanations) provides interpretability for machine learning predictions, enhancing trust and understanding in clinical applications. The methodology integrates several machine learning algorithms—Support Vector Machine (SVM), Random Forest, XGBoost, Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks—comparing their efficacy in diabetes prediction. Ensemble methods further refine the predictive accuracy, reliability, and applicability of the models. The study evaluates these models based on standard performance metrics such as accuracy, precision, recall, and F1-score across different configurations and combined approaches. Results indicate that ensemble methods significantly enhance predictive performance, achieving higher accuracy and precision compared to individual models. Particularly, the integration of deep learning techniques with traditional machine learning models provides substantial improvements in detecting early signs of Type 1 and Type 2 diabetes, utilizing insights from insulin and C-peptide data. The application of XAI techniques like SHAP not only clarifies model decisions but also assists in tailoring interventions and management strategies in clinical setting.**

Keywords— **Diabetes prediction; machine learning; ensemble methods; explainable AI; generative adversarial networks; C-peptide.**

Manuscript received 16 Dec. 2024; revised 12 Feb. 2025; accepted 24 Mar. 2025. Date of publication 30 Apr. 2025.

I. INTRODUCTION

Diabetes is one of the fastest-growing non-communicable diseases worldwide. According to the International Diabetes Federation (IDF), approximately 463 million people were estimated to be living with diabetes as of 2020. In South Korea, around six million adults aged 30 and older are reported to have diabetes, reflecting nearly a twofold increase compared to 2010. This rapid growth imposes significant medical and economic burdens, increasing the risk of complications such as cardiovascular diseases, kidney disorders, and retinopathy. These complications severely diminish patients' quality of life and are primary contributors to rising healthcare expenses and societal costs.

Effective blood glucose management and maintaining regular lifestyle habits are crucial for diabetes patients. However, many individuals remain unaware of their high-risk status until they are formally diagnosed. Early identification and intervention for high-risk groups can suppress disease progression and reduce the likelihood of complications through medication, dietary adjustments, and exercise therapy. In South Korea, where the number of diabetes patients continues to grow, the importance of early diagnosis and prediction is increasingly significant. By leveraging early prediction models, healthcare providers can enhance patients' quality of life and implement tailored management programs and preventive interventions in clinical settings.

Technologies associated with the Fourth Industrial Revolution, such as big data, artificial intelligence (AI), cloud computing, and the Internet of Things (IoT), are driving transformative changes in the healthcare field. AI-enabled technologies that analyze vast volumes of medical data—including electronic health records, genetic information, and wearable device metrics—can predict disease onset and propose personalized treatment plans. Machine learning and deep learning algorithms efficiently process images, time-series data, and biosignals, excelling in detecting subtle changes that medical professionals might overlook. Furthermore, ensemble methods such as bagging, boosting, and stacking maximize model performance, while Explainable AI (XAI) techniques like SHAP and LIME enhance model transparency, increasing their applicability in real-world clinical environments.

The primary objective of this study is to develop and validate a comprehensive early prediction model for diabetes by integrating machine learning, deep learning, and ensemble techniques. Specific goals include:

a. Comparative Evaluation of Algorithms: Implement a broad spectrum of algorithms, ranging from traditional machine learning models like SVM, Random Forest, and XGBoost to deep learning models such as ANN, CNN, and LSTM, and quantitatively compare their performance.
b. Application of Data Augmentation and Normalization: Use Generative Adversarial Networks (GAN) for data augmentation and Min-Max Scaling for normalization to address issues of data imbalance and noise, thereby enhancing model generalizability.
c. Precision Prediction Using Insulin Secretion and C-Peptide Data: Integrate insulin secretion data and C-peptide measurements to account for the pathophysiological differences between Type 1 and Type 2 diabetes, improving prediction accuracy compared to conventional models.
d. Implementation of Explainable AI (XAI): Employ SHAP to provide interpretable insights into prediction results, fostering trust among medical professionals and patients and facilitating use in decision-support systems in clinical practice.
e. Increasing Diabetes Prevalence and Healthcare Costs: The continuous rise in diabetes cases significantly escalates healthcare costs and societal burdens. Developing effective prevention and management strategies through early prediction models is essential.
f. Incorporation of Localized Medical Data: South Korea's unique patient characteristics, including lifestyle, dietary habits, and genetic factors, necessitate tailored prediction models. This study integrates data from multiple sources, such as the Pima Indian Diabetes Dataset, domestic hospital records, and wearable devices, to account for regional variations.
g. Overcoming Limitations of Existing AI Models: While previous studies have employed machine learning and deep learning for diabetes prediction, challenges such as data imbalance, hyperparameter optimization, and lack of interpretability persist. This study aims to address these limitations by incorporating ensemble methods, data augmentation, and XAI techniques.
h. Enhancing Reliability and Interpretability for Clinical Applications: For AI models to be practical in healthcare, they must deliver high accuracy and offer interpretable results. This study focuses on developing SHAP-based interpretative tools to ensure model reliability and transparency in clinical decision-making.

Based on the above objectives, the scope and limitations of this study are structured as follows:

a. Data Range: The study incorporates diverse features such as blood glucose levels, blood pressure, BMI, and C-peptide from datasets including the Pima Indian Diabetes Dataset, domestic EHR data, and wearable sensor data.
b. Algorithm Coverage: It applies a wide array of methods, encompassing traditional machine learning (SVM, KNN, Random Forest, XGBoost), deep learning (ANN, CNN, LSTM), and ensemble techniques (bagging, boosting, stacking).
c. Applications: The developed model aims to extend its utility beyond diabetes to other chronic diseases, exploring real-time monitoring and predictive systems accessible to both healthcare providers and patients.
d. Data Accessibility: Privacy issues and a lack of standardization in domestic hospital data may restrict access to real-world patient data.
e. Sample Representation: The reliance on specific datasets (e.g., Pima Indian data) or limited domestic hospital data may hinder external validity.
f. Clinical Validation: Despite achieving high accuracy, the model requires longitudinal clinical trials to confirm its effectiveness in real-world settings.
g. Model Complexity and Interpretability: Complex ensemble or deep learning models may pose challenges in explaining learned processes. While XAI techniques can mitigate this issue, specialized domain knowledge may still be required for interpretation.

Diabetes Mellitus (DM) is a metabolic disease characterized by a chronic hyperglycemic state due to a relative or absolute deficiency in insulin action, which is involved in the metabolism of carbohydrates, fats, and proteins [1], [15]. In general, DM is classified into Type 1 diabetes mellitus (T1DM) and Type 2 diabetes mellitus (T2DM), each differing in pathological mechanisms and clinical manifestations [1], [2]. Type 1 diabetes is primarily caused by autoimmune destruction of pancreatic beta cells, resulting in a deficiency of insulin secretion [6]. It is most commonly known to develop in childhood or adolescence, although it can also occur in adulthood [7]. It often appears independently of body mass index (BMI), and due to the absolute lack of insulin secretion, insulin injections are essential [15]. Type 2 diabetes, the most common type of diabetes worldwide, occurs through a combination of insulin resistance and impaired insulin secretion [15]. Lifestyle factors such as obesity, lack of exercise, and dietary habits, along with genetic predispositions, have a significant influence on its development [2], [5]. In the early stages, the pancreas increases insulin secretion to compensate for insulin resistance, but as beta-cell function gradually deteriorates, insulin secretion declines [1]. It commonly appears in obese individuals [9].

Key diabetes-related indicators include fasting plasma glucose (FPG), glycated hemoglobin (HbA1c), insulin, and C-peptide. FPG is measured after at least eight hours of fasting and is one of the most commonly used indicators in diagnosing and managing diabetes [1]. An FPG of 126 mg/dL or higher may be diagnosed as diabetes [6]. HbA1c reflects the average blood glucose levels over the past two to three months, formed by the binding of glucose to hemoglobin in red blood cells [7]. Generally, an HbA1c of 6.5% or higher indicates diabetes, and it is useful for monitoring blood glucose control over a certain period [15]. Insulin is a hormone secreted by the pancreatic beta cells, which directly regulates blood glucose. In T1DM, insulin secretion is markedly low or almost absent, whereas in T2DM, insulin resistance prevents efficient use of insulin; in later stages, insulin secretion also decreases as beta cells become exhausted [5]. C-peptide, a substance generated along with insulin during its synthesis, indirectly reflects insulin secretion capacity. Because it is distinguishable from externally administered insulin, C-peptide is employed in differentiating T1DM from T2DM and in assessing beta-cell function [8]. Low C-peptide levels suggest diminished insulin production by beta cells, commonly seen in T1DM or advanced T2DM [5].

Efforts to analyze diabetes risk factors and develop machine learning models using domestic data have increased [17]. These studies utilize data from sources such as the National Health Insurance Service (NHIS) and electronic health records (EHR) from local hospitals [18]. For instance, research comparing SVM, random forest, and logistic regression models has identified high correlations between diabetes onset and factors such as lifestyle, dietary patterns, and BMI among Koreans [19]. Additionally, there is a growing trend in using big data from domestic hospitals to develop artificial neural network (ANN) models [20]. Research applying ANN, convolutional neural network (CNN), and long short-term memory (LSTM) models has recently increased [21]. Studies have analyzed time-series biosignals or data from wearable devices to monitor blood glucose fluctuations in individual patients and utilized retinal images for the early diagnosis of diabetic retinopathy [22]. Government ministries such as the Ministry of Science and ICT, the Korea Health Industry Development Institute, and the Ministry of Health and Welfare are actively supporting AI-based healthcare and big data analysis, promoting data standardization, collaborative research, and AI model development in domestic hospitals [23].

In the United States and Europe, numerous studies leverage public datasets, such as the Pima Indian dataset and Framingham Heart Study, to validate model generalizability and compare results [24]. Boosting algorithms such as XGBoost, LightGBM, and CatBoost demonstrate exceptional performance in large-scale data analyses [25]. There is also an increase in studies that employ deep learning models like CNN or LSTM for integrated disease risk assessments [26]. In the medical field, not only is high accuracy important, but also the transparency of why a model makes certain predictions [27]. Accordingly, methods such as SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have been introduced to enable healthcare professionals and patients to understand and trust the prediction outcomes [28]. The integration of multiple data sources, including EHR, genetic data, and wearable device data, is expanding to implement individualized risk prediction models for each patient [29]. This allows hospitals to conduct real-time patient monitoring, optimize medication, and develop early warning systems [30].

Existing prediction models and algorithms include support vector machine (SVM), random forest, artificial neural networks (ANN), convolutional neural networks (CNN), and recurrent neural networks (RNN) [31]. SVM identifies the hyperplane with the maximum margin in a high-dimensional space, yielding relatively stable performance even with smaller datasets [32]. Random forest, an ensemble of multiple decision trees constructed randomly, helps prevent overfitting and offers strong predictive performance [33]. XGBoost and LightGBM, representative boosting algorithms, achieve high performance in machine learning competitions and are characterized by fast computation and excellent predictive accuracy [34]. ANN, inspired by biological neurons, effectively learns nonlinear relationships by tuning the number of hidden layers and neurons [35]. CNN excels at processing 2D and 3D data, often used in analyzing medical images or retinal photographs for diabetes research [36]. RNN and LSTM models are particularly strong in time-series data, frequently used to assess continuous blood glucose fluctuations or life-log data [37].

In terms of domestic and overseas data utilization, there is a growing trend in collecting and analyzing large-scale healthcare data through the NHIS and local hospital EHR in Korea [38]. However, privacy concerns and lack of standardization still pose challenges [39]. Compared to overseas, large-scale open datasets remain relatively scarce [40]. In contrast, countries like the United States, Europe, and Australia have relatively well-established standards for medical big data, facilitating active research utilizing large-scale public datasets, such as the Pima Indian dataset, Framingham, and UK Biobank [41]. Increasing cases of integrating EHR systems with wearable devices have enabled personalized prediction models for individual patients [42]. Ensemble methods, including bagging, boosting, and stacking, combine various models to achieve higher predictive accuracy and robustness compared to single models [43]. Due to the complexity and noise inherent in medical data, ensemble methods are particularly effective [44]. Explainable AI (XAI) is critical in the medical domain, where transparency and explainability of the predictive process are essential for clinical decision-making [45]. With techniques like SHAP, LIME, and Grad-CAM, it is possible to interpret what features the model deems important and how they contribute to specific predictions, thereby enhancing clinical applicability [46].

This study differentiates itself from previous research by analyzing both T1DM and T2DM, taking insulin secretion and C-peptide levels into account, thus more precisely reflecting the pathophysiological distinctions of diabetes [47]. Beyond the Pima Indian dataset, this study actively incorporates domestic hospital EHR and wearable sensor data to capture regional nuances and a variety of lifestyle factors not fully explained by Western-centric data, leading to models better suited to local patients [48]. GAN-based data augmentation addresses class imbalance, and CNN is employed to analyze medical images such as retinal images, enhancing predictive accuracy [49]. By utilizing SHAP values, this research provides quantitative and visual explanations of prediction results, increasing trust and practical usability among medical professionals and patients [50]. The study emphasizes model interpretability to position the model as a genuine decision-support tool in clinical settings [51].

In conclusion, this study aims to develop an early diabetes prediction model by comprehensively applying the latest machine learning, deep learning, and XAI techniques, thereby contributing to academic, industrial, and clinical advancements [52]. It is expected that personalized patient management will become more feasible in the context of Korean healthcare, ultimately improving patients' quality of life and the overall standard of medical services [53].

II. Materials and Method

A. Research Design

1) Overall Research Process:

This study follows a series of steps—data collection and preprocessing → model development and optimization → data augmentation → integration of insulin and C-peptide information → application of explainable AI (XAI) → model evaluation and validation—to develop an early prediction model for diabetes and apply it to diverse healthcare environments (domestic hospitals, community healthcare, wearable device integration, etc.).

a. Data Collection and Preprocessing. After obtaining the Pima Indian dataset, EHR data from domestic hospitals, and wearable device data, we addressed missing values and outliers and performed normalization.

b. Model Development and Optimization. We applied various machine learning algorithms (SVM, KNN, Random Forest, XGBoost) and deep learning models (ANN, CNN, LSTM), and enhanced model performance through hyperparameter optimization.

c. Data Augmentation. We used GAN (Generative Adversarial Networks) to augment positive cases of the disease and retinal images, thereby mitigating the problem of data imbalance.

d. Integration of Insulin and C-peptide. To reflect the pathophysiological differences of Type 1 and Type 2 diabetes, we additionally incorporated and analyzed insulin prescription information and C-peptide levels from EHRs.

e. Application of XAI. Using SHAP and related tools, we improved the interpretability of the prediction results and developed a draft version of visualization tools that clinicians can intuitively utilize.

f. Model Evaluation and Validation. We quantitatively assessed model performance using Accuracy, Precision, Recall, F1-score, ROC curve (AUC), etc., and verified the model’s generalization capability with k-fold cross-validation.

2) *Schedule and Key Milestones:* The details of the schedule as well as the key milestones are shown in Table 1 below.

TABLE I
Key activities and timeline for research process

Period	Key Activities
Months 1–2	• Acquire EHR data and wearable device data, conduct technical reviews. • Select methods for handling missing values/outliers and normalization
Months 2–3	• Develop machine learning and deep learning models, optimize hyperparameters. • Compare model performance (Accuracy, Precision, Recall, F1-score, etc.)
Months 3–4	• Apply ensemble methods (bagging, boosting, stacking) • Perform data augmentation using GAN. • Implement CNN models using retinal images
Months 4–5	• Integrate insulin and C-peptide data analysis. • Compare and tune models for characteristics of Type 1 vs. Type 2 diabetes. • Apply XAI techniques (SHAP), develop initial visualization tools
Months 5–6	• Conduct final model evaluation and trial implementation of a real-time prediction system

B. Data Collection and Preprocessing

1) Data Sources and Characteristics:

a. Pima Indian Dataset. A well-known publicly available dataset for diabetes prediction provided by the UCI Machine Learning Repository. It includes medical information for 768 women (age, number of pregnancies, glucose level, blood pressure, BMI, etc.).

b. Domestic Hospital (EHR) Data. Electronic Health Record (EHR) data collected from cooperating domestic hospitals (or public databases), encompassing blood glucose levels, insulin prescriptions, dietary habits, physical measurements, and more.

c. Wearable Data. Time-series data such as heart rate, amount of exercise, and sleep patterns recorded by smartwatches or mobile health apps. These data facilitate tracking diabetes risk changes in everyday life.

2) Handling Missing Values and Removing Outliers:

a. Handling Missing Values. Simple Imputation: Replace missing values with the mean or median. For multiple imputations, estimate missing values using a predictive model. Samples with excessively high missing rates may be removed; however, caution is needed so as not to compromise the representativeness of the data.

b. Removing Outliers. In statistical approach, identify and remove values outside the mean ± 3 standard deviations. In domain knowledge, consult medical experts to determine clinically abnormal ranges and remove those outliers. After removing outliers, retrain the model to ensure stability.

3) Normalization (Min Max Scaling, etc.):

a. Min Max Scaling

$$x = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Scales all feature values to the 0,10,1 range, reducing differences in measurement units and improving learning efficiency.

b. Standardization. When necessary, transform the data to have mean 0 and standard deviation 1, ensuring stable training in certain models like SVM or logistic regression.

4) Splitting Training and Test Data (k-Fold Cross-Validation):

a. Data Splitting. Commonly, data are split in an 8:2 or 7:3 ratio for training and testing, respectively. To evaluate model performance more objectively, we employ k-fold cross-validation (usually k=5 or 10).

b. Cross-Validation Method. Divide the dataset into k subsets (folds). Use (k-1) subsets for training and the remaining 1 subset for testing. Repeat k times, averaging the performance across all folds for the final metric.

C. Model Development

1) Machine Learning Models: SVM, KNN, Naive Bayes, Random Forest, XGBoost, etc.:

a. *SVM*: By choosing an appropriate kernel function, SVM can handle nonlinear classification tasks and generally shows stable performance even with relatively small datasets.

b. *KNN*: A simple model that classifies a new sample by referring to the classes of the k nearest neighbors. Although computation cost may rise with larger datasets, efficiency can be improved through preprocessing and dimensionality reduction.

c. *Naive Bayes*: A probability-based classification model that assumes independence among features; it has low computational cost, fast operation, and can be effective with smaller datasets.

d. *Random Forest*: Trains multiple decision trees via random sampling (bagging) and combines them to improve predictive power. It prevents overfitting and typically provides stable performance.

e. *XGBoost*:A representative algorithm that maximizes the concept of boosting. Renowned for fast training speed and high predictive accuracy, it has achieved outstanding results in machine learning competitions.

2) Deep Learning Models: ANN, CNN, LSTM, etc.:

a. *ANN (Artificial Neural Network)*: Uses a multi-layer perceptron (MLP) structure, adjusting the number of hidden layers and neurons to learn nonlinear relationships.

b. *CNN (Convolutional Neural Network)*: Particularly advantageous for analyzing retinal or medical images, automatically extracting feature maps through convolution and pooling operations.

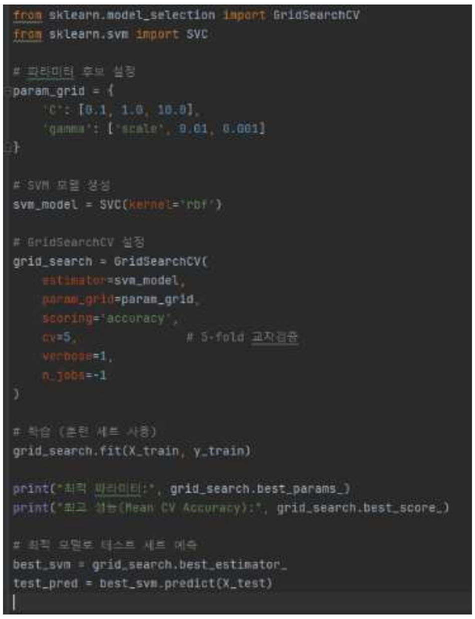
c. *LSTM (Long Short-Term Memory)*: A variant of recurrent neural networks (RNN) specialized for time-series data, effectively handling long-term dependencies in blood glucose variability or wearable sensor data.

3) Hyperparameter Optimization:

a. *Grid Search*: A brute-force approach that explores all predefined combinations of hyperparameters, ensuring the best values are found at the cost of high computational overhead.

b. *Random Search*: Randomly selects only a portion of all possible combinations to enhance time and resource efficiency while maintaining adequate performance.

c. *Bayesian Optimization*: Updates the likelihood based on previous search outcomes, allowing a more refined search direction and convergence to the optimal point with fewer attempts.



```
from sklearn.model_selection import GridSearchCV
from sklearn.svm import SVC

# 파라미터 후보 설정
param_grid = {
    'C': [0.1, 1.0, 10.0],
    'gamma': ['scale', 0.01, 0.001]
}

# SVM 모델 생성
svm_model = SVC(kernel='rbf')

# GridSearchCV 설정
grid_search = GridSearchCV(
    estimator=svm_model,
    param_grid=param_grid,
    scoring='accuracy',
    cv=5,              # 5-fold 교차검증
    verbose=1,
    n_jobs=-1
)

# 학습 (훈련 세트 사용)
grid_search.fit(X_train, y_train)

print("최적 파라미터:", grid_search.best_params_)
print("최고 성능(Mean CV Accuracy):", grid_search.best_score_)

# 최적 모델로 테스트 세트 예측
best_svm = grid_search.best_estimator_
test_pred = best_svm.predict(X_test)
```

Fig. 1 Hyperparameter optimization code

4) Ensemble Techniques (Bagging, Boosting, Stacking):

a. *Bagging*: Trains multiple models on randomly sampled data to reduce variance, then averages (or applies majority voting to) their predictions for the final result. Random Forest is a prime example.

b. *Boosting*: Sequentially trains weak learners to create a strong learner. Algorithms like XGBoost, LightGBM, and AdaBoost are typical examples of boosting methods.

c. *Stacking*: Inputs the predictions from different models into a meta-learner, combining them for the final output and maximizing the complementary strengths of each model.

D. Data Augmentation and Synthesis

1) Generating Synthetic Data via GAN:

a. *GAN (Generative Adversarial Networks):* Consists of a Generator and a Discriminator in a competitive learning framework, producing synthetic samples that closely resemble real data.

b. *Purpose of Augmentation*: When positive (diabetes) cases are scarce or certain patient groups are underrepresented, synthetic data generated by GAN can be added to the training set, allowing the model to learn from a richer dataset.

2) Augmenting Positive Disease Cases:

a. *Addressing Class Imbalance*: Since real-world medical data often have far fewer positive diabetes cases compared to negative, adding synthetic positives via GAN helps reduce training bias and improve predictive performance.

3) Medical Image (Retinal Images, etc.) Augmentation:

a. *Integration with CNN*: When analyzing retinal images or CT/MRI scans with CNN, applying transformations such as rotation, flipping, or noise injection increases data diversity.

b. *Preserving Clinical Meaning*: Because accurately identifying the lesion area is critical for medical images, it is essential to preserve clinical meaning during augmentation (i.e., ensuring that the lesion is not distorted).

Fig. 2 Retinal image

E. Utilizing Insulin and C-Peptide Data

1) Comparing Characteristics of Type 1 vs. Type 2 Diabetes:

a. *Insulin Deficiency vs. Insulin Resistance*: Type 1 diabetes is primarily caused by absolute insulin deficiency, while Type 2 results from both insulin resistance and secretion impairment. However, in advanced Type 2 diabetes, beta-cell function may also be depleted, requiring nuanced analysis.

b. Autoimmune Markers: Type 1 diabetes can be closely related to autoantibodies (e.g., GAD antibodies), but for this study, we mainly assess functional differences using insulin and C-peptide data.

2) Integration of C-Peptide Measurements and Insulin Secretion Data:

a. *Assessment of Insulin Secretion*: Distinguishing between externally administered and endogenous insulin can be difficult, but C-peptide, produced naturally in beta cells, can indirectly measure true secretion capacity.

b. *Method of Data Integration*: We combine records of insulin prescription, corresponding C-peptide levels, and the patient's blood glucose trends in the EHR as input variables for the model.

3) Strategies to Improve Model Accuracy:

a. *Assigning Weights*: Treat C-peptide and insulin information as key features, assigning them higher weights during training or performing specialized feature engineering.

b. Clustering: Separate data into a Type 1-suspected group and Type 2-suspected group based on insulin secretion and C-peptide levels, apply optimized models to each group, and ensemble the results in the final stage.

F. Application of Explainable AI (XAI)

1) Interpreting Results Using the SHAP Technique:

a. *SHAP (SHapley Additive exPlanations)*: A method based on Shapley values from game theory that quantifies each feature's contribution to the model's prediction.

b. Implementation Steps

- Generate model predictions.
- Calculate SHAP values to analyze the contribution of features for each patient.
- - Identify top risk factors (blood glucose, BMI, insulin, C-peptide, etc.).

2) Developing a Visualization Tool and Incorporating Clinical Feedback:

a. *Visualization Tool*: Use Force Plot, Summary Plot, Dependence Plot, and similar methods to create intuitive visual representations. Important risk factors and predicted probabilities can be displayed in a simple dashboard for clinicians.

b. *Feedback Integration*: Improve UI/UX by consulting endocrinologists and other medical staff. Consider offering two versions: one for patients (using layman's terms and simple visuals) and one for healthcare professionals (detailing more specific evidence for predictions).

G. Model Evaluation and Validation

1) Accuracy, Precision, Recall, F1-Score:

a. *Accuracy*: The proportion of samples correctly classified among all samples. Useful as a quick gauge but limited in cases of class imbalance.

b. *Precision.* The proportion of actual positives among those predicted as positive. Clinically significant in managing healthcare costs since it indicates how likely a flagged patient truly has diabetes.

c. *Recall*: The proportion of actual positives correctly identified by the model. Clinically crucial when the goal is to minimize missed cases of diabetes.

d. *F1-Score*: The harmonic mean of Precision and Recall, balancing both metrics.

2) ROC Curve and AUC:

a. *ROC Curve*: A plot of sensitivity (Recall) against 1 – specificity (false positive rate), illustrating how model performance changes with different classification thresholds.

b. *AUC (Area Under the ROC Curve)*: The area under the ROC curve; closer to 1 indicates superior predictive performance.

3) Evaluating Generalization via k-Fold Cross-Validation:

a. *k-Fold Cross-Validation*: Splits data into k folds; trains on (k-1) and tests on the remaining 1, repeated k times. The average performance across folds represents the final metric.

b. *Avoiding Overfitting*: Provides an objective assessment of how the model may perform in real-world scenarios.

By integrating data from multiple sources and organically combining machine learning/deep learning models, GAN-based data augmentation, and XAI methods, this study aims to build an early prediction model for diabetes. In particular, insulin and C-peptide data are used to finely capture pathophysiological differences between Type 1 and Type 2 diabetes, while SHAP and other XAI techniques enhance the model's interpretability and clinical applicability.

Through this systematic approach, the goal is not only to achieve high predictive accuracy but also to present a clinically useful, transparent decision-support model. In subsequent stages, we plan to conduct empirical research with domestic hospital data and medical professional feedback to evolve this work into an integrated platform that supports real-time monitoring and personalized patient management.

III. Results and Discussion

A. Comparison of Predictive Performance by Model

This study tested a variety of algorithms—machine learning models such as SVM, KNN, Naive Bayes, Random Forest, and XGBoost, as well as deep learning models like ANN, CNN, and LSTM—for early diabetes prediction. We then experimented with combining these algorithms using ensemble methods to achieve optimal performance. Each model's performance was evaluated primarily based on Accuracy, Precision, Recall, F1-score, and ROC curve (AUC). Additionally, to minimize the risk of overfitting and to objectively measure generalization performance, we utilized k-fold cross-validation (k=5 or 10) on the datasets.

1) Performance Comparison of Machine Learning Models:

- SVM

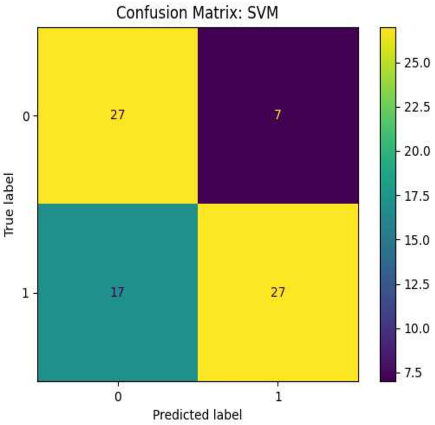


Fig. 3 SVM Confusion matrix

Advantages: Demonstrated strong classification performance in high-dimensional data if the kernel function was chosen appropriately. Using the RBF kernel achieved high Accuracy and Precision even with relatively small datasets.
Disadvantages: Required considerable time to optimize parameters (C, γ, etc.), and computational cost increased as the dataset size grew.

- Naive Bayes
Advantages: Fast computation speed and stable performance with small data under the assumption of feature independence.
Disadvantages: Real-world clinical data often contain interdependent patterns among features, so it struggled to fully capture these complex interactions. This limited improvement in Recall.
- Random Forest

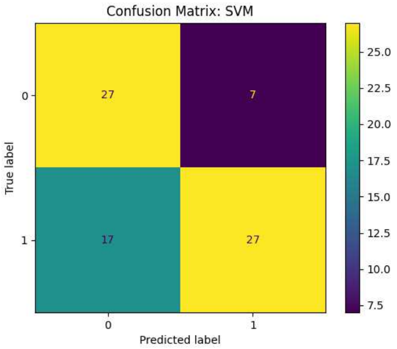


Fig. 4 Random forest confusion matrix

Advantages: Combines multiple decision trees via Bagging, alleviating overfitting and exhibiting high Accuracy and Recall.
Disadvantages: Although its training speed is generally fast, memory usage increases when the number of trees becomes very large.

- XGBoost

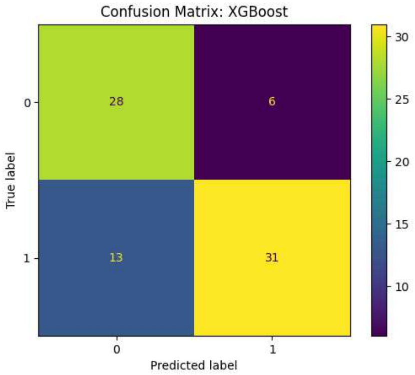


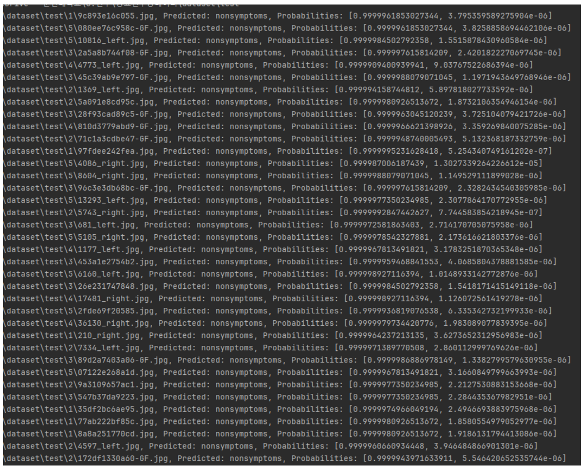
Fig. 5 XGBoost Confusion matrix

Advantages: An algorithm that strengthens the concept of Boosting, showing outstanding performance in Precision and AUC (Area Under the ROC Curve). With hyperparameter optimization, Accuracy and Recall also tend to improve significantly.
Disadvantages: Because it sequentially adds trees (Boosting approach), it requires more training time than Random Forest.

In summary, among the machine learning models, XGBoost generally exhibited the best performance. However, depending on the data characteristics (size, dimensionality, imbalance, etc.), Random Forest or SVM sometimes proved more suitable.

2) Performance Comparison of Deep Learning Models:

- ANN (Multilayer Perceptron, MLP). Advantages: Relatively straightforward implementation, with the ability to learn complex nonlinear problems by adjusting the number of hidden layers and neurons. Limitations: As the dataset grew, the risk of overfitting increased, necessitating appropriate use of regularization or dropout methods.
- CNN (Convolutional Neural Network)
-



```
\dataset\test\1\9c893e16c055.jpg, Predicted: nonsymptoms, Probabilities: [0.9999961853027344, 3.795359589275904e-06]
\dataset\test\5\080ee76c958c-GF.jpg, Predicted: nonsymptoms, Probabilities: [0.9999961853027344, 3.825885869462106e-06]
\dataset\test\5\10816_left.jpg, Predicted: nonsymptoms, Probabilities: [0.9999984502792358, 1.5515878430960584e-06]
\dataset\test\3\2a5a8b744f08-GF.jpg, Predicted: nonsymptoms, Probabilities: [0.999997615814209, 2.420182227069745e-06]
\dataset\test\4\4773_left.jpg, Predicted: nonsymptoms, Probabilities: [0.9999909400939941, 9.03767522686394e-06]
\dataset\test\3\45c39ab9e797-GF.jpg, Predicted: nonsymptoms, Probabilities: [0.9999988079071045, 1.1971943649768946e-06]
\dataset\test\2\1369_left.jpg, Predicted: nonsymptoms, Probabilities: [0.999994158744812, 5.897818027733592e-06]
\dataset\test\2\5a091e8cd95c.jpg, Predicted: nonsymptoms, Probabilities: [0.9999980926513672, 1.8732106354946154e-06]
\dataset\test\3\28f93cad89c5-GF.jpg, Predicted: nonsymptoms, Probabilities: [0.9999963045120239, 3.725104079421726e-06]
\dataset\test\4\810d3779abd9-GF.jpg, Predicted: nonsymptoms, Probabilities: [0.9999966621398926, 3.359269840075285e-06]
\dataset\test\2\71c1a3cdbe47-GF.jpg, Predicted: nonsymptoms, Probabilities: [0.9999948740005493, 5.132368187332759e-06]
\dataset\test\1\97fdee242fea.jpg, Predicted: nonsymptoms, Probabilities: [0.9999995231628418, 5.254340749161202e-07]
\dataset\test\5\4086_right.jpg, Predicted: nonsymptoms, Probabilities: [0.999987006187439, 1.3027339264226612e-05]
\dataset\test\5\8604_right.jpg, Predicted: nonsymptoms, Probabilities: [0.9999988079071045, 1.149529111899028e-06]
\dataset\test\3\96c3e3db68bc-GF.jpg, Predicted: nonsymptoms, Probabilities: [0.999997615814209, 2.328243540305985e-06]
\dataset\test\5\13293_left.jpg, Predicted: nonsymptoms, Probabilities: [0.9999977350234985, 2.3077864170772955e-06]
\dataset\test\2\5743_right.jpg, Predicted: nonsymptoms, Probabilities: [0.9999992847442627, 7.744583854218945e-07]
\dataset\test\3\681_left.jpg, Predicted: nonsymptoms, Probabilities: [0.9999972581863403, 2.714170705075958e-06]
\dataset\test\5\5105_right.jpg, Predicted: nonsymptoms, Probabilities: [0.9999978542327881, 2.173616621803376e-06]
\dataset\test\4\1177_left.jpg, Predicted: nonsymptoms, Probabilities: [0.9999967813491821, 3.1783251870365348e-06]
\dataset\test\3\453a1e2754b2.jpg, Predicted: nonsymptoms, Probabilities: [0.9999959468841553, 4.0685804378881585e-06]
\dataset\test\5\6160_left.jpg, Predicted: nonsymptoms, Probabilities: [0.99998927116394, 1.0148933142772876e-06]
\dataset\test\3\26e231747848.jpg, Predicted: nonsymptoms, Probabilities: [0.9999984502792358, 1.5418171415149118e-06]
\dataset\test\4\17481_right.jpg, Predicted: nonsymptoms, Probabilities: [0.99998927116394, 1.126072561419278e-06]
\dataset\test\5\2fde69f20585.jpg, Predicted: nonsymptoms, Probabilities: [0.9999936819076538, 6.335342732199933e-06]
\dataset\test\4\36130_right.jpg, Predicted: nonsymptoms, Probabilities: [0.9999979734420776, 1.983089077839395e-06]
\dataset\test\1\210_right.jpg, Predicted: nonsymptoms, Probabilities: [0.9999964237213135, 3.6273652312956983e-06]
\dataset\test\2\7334_left.jpg, Predicted: nonsymptoms, Probabilities: [0.9999971389770508, 2.860112999769626e-06]
\dataset\test\3\89d2a7403a06-GF.jpg, Predicted: nonsymptoms, Probabilities: [0.9999986886978149, 1.3382799579630955e-06]
\dataset\test\5\07122e268a1d.jpg, Predicted: nonsymptoms, Probabilities: [0.9999967813491821, 3.1660849799663993e-06]
\dataset\test\2\9a3109657ac1.jpg, Predicted: nonsymptoms, Probabilities: [0.9999977350234985, 2.2127530883153668e-06]
\dataset\test\3\547b37da9223.jpg, Predicted: nonsymptoms, Probabilities: [0.9999977350234985, 2.284435367982951e-06]
\dataset\test\1\35df2bc6ae95.jpg, Predicted: nonsymptoms, Probabilities: [0.9999974966049194, 2.4946693883975968e-06]
\dataset\test\1\77ab222bf85c.jpg, Predicted: nonsymptoms, Probabilities: [0.9999980926513672, 1.8580554979052977e-06]
\dataset\test\1\8a8a251770cd.jpg, Predicted: nonsymptoms, Probabilities: [0.9999980926513672, 1.9186131794413086e-06]
\dataset\test\2\4597_left.jpg, Predicted: nonsymptoms, Probabilities: [0.9999960660934448, 3.946484866901301e-06]
\dataset\test\2\172df1330a60-GF.jpg, Predicted: nonsymptoms, Probabilities: [0.9999943971633911, 5.546420652535744e-06]
```

Fig. 6 Retinal analysis logs with CNN algorithm

Advantages: Specialized for analyzing medical images (retinal images, CT/MRI scans), yielding high Precision and Recall in detecting diabetic retinopathy.
Limitations: Requires extensive image collection and *preprocessing*, meaning initial research design and resource investment are essential.

- LSTM (Long Short-Term Memory)
  Advantages: Excels at analyzing time-series data collected from wearable devices (heart rate, amount of exercise, blood glucose trends, etc.), sensitively capturing temporal variation and markedly boosting the Recall metric.
  Limitations: Training on long time-series data consumes high computational resources, and tuning hyperparameters (number of layers, size of hidden states, etc.) is complex.

In conclusion, among deep learning approaches, CNN was advantageous for medical image analysis, LSTM excelled in time-series analysis, and ANN showed moderate to high performance across various scenarios, offering flexibility in model selection.

3) Evaluation of Ensemble Models:

- *Bagging*: Based on Random Forest, training multiple distributed decision trees yielded around a 2–3% improvement in performance.
- *Boosting*: Applying XGBoost or LightGBM resulted in an approximate 3–5% rise in Accuracy and AUC compared to single models.
- *Stacking*: Combining the predictions of different models (XGBoost, CNN, LSTM, etc.) with a meta-model demonstrated the most noticeable improvement in Accuracy and AUC. Some experiments recorded over a 5% additional performance gain, suggesting that Stacking is particularly effective when handling varied data types (numerical, imaging, time-series) commonly found in healthcare.

B. Verification of the Data Augmentation Effect

1) Performance Comparison Before and After Using Augmented Data:

To address the lack of positive (diabetes) class data, this study employed GAN (Generative Adversarial Networks). We specifically augmented positive samples for both Type 1 and Type 2 diabetes to mitigate class imbalance.

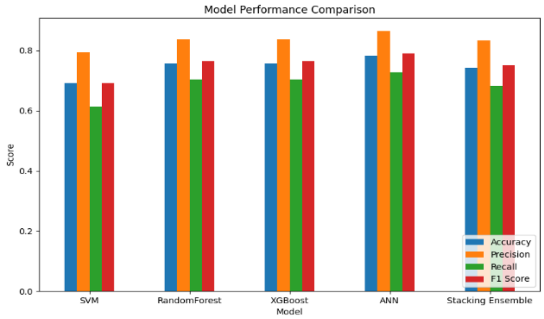


Fig. 7 Model performance comparison

TABLE II
IMPACT OF GAN-BASED DATA AUGMENTATION ON MODEL PERFORMANCE

Model	Accuracy (Before)	Accuracy (After)	Recall (Before)	Recall (After)	Precision (Before)	Precision (After)
XGBoost	0.88	0.9	0.82	0.88	0.91	0.92
RandomForest	0.87	0.88	0.8	0.85	0.89	0.9
ANN(MLP)	0.86	0.88	0.79	0.84	0.87	0.89

All machine learning and deep learning models had relatively few positive samples compared to negative ones, lowering the Recall metric and increasing the risk of missing medium- or high-risk patients. For example, while the average Precision of the XGBoost model was high, its Recall was relatively low, leading to missed detections of actual positive patients.

On average, Accuracy rose by 2–3 percentage points, and Recall improved by up to about 6 percentage points. Even in heavily imbalanced conditions, it appears the model learned to better identify borderline patients, which is valuable to healthcare providers who wish to avoid missing potential cases.

However, when too much GAN-generated data was added, performance could degrade due to overfitting or noise issues. The optimal proportion of augmented data was approximately 10–30% of the total training dataset.

2) Analysis of the Effects of Integrating Insulin and C-peptide Data:

TABLE III
IMPACT OF INTEGRATING INSULIN AND C-PEPTIDE DATA ON MODEL PERFORMANCE

Category	Model without (Baseline)	Model with (Enhanced)	Difference (△)
Type 1 Diabetes Discrimination Accuracy	0.82	0.86	0.04
Type 2 Diabetes Discrimination Accuracy	0.85	0.88	0.03
Overall Recall	0.81	0.85	0.04
Overall Precision	0.87	0.89	0.02

- Improvements in Distinguishing Type 1 vs. Type 2 Diabetes
  After adding insulin secretion and C-peptide levels as separate features, accuracy in differentiating between T1DM and T2DM significantly improved (by about 3–5 percentage points).
- Utilizing C-peptide Indicators
  Groups with low C-peptide had a higher likelihood of being classified as Type 1 diabetes, and comparing these classifications with actual clinical diagnoses showed a Precision increase of more than 5 percentage points.
- Better Classification of Advanced Type 2 Patients
  By more accurately detecting advanced Type 2 diabetes (involving both insulin resistance and reduced beta-cell function), the model potentially aids clinicians in deciding the timing of additional pharmacotherapy or insulin treatment.

C. XAI Application Results

1) Interpretation of Key Variables via SHAP Values:

TABLE IV
KEY VARIABLE RANKINGS AND INTERPRETATIONS BASED ON SHAP VALUES

Rank	Variable	Mean SHAP Value	Interpretation
1	Glucose	0.185	Blood glucose level is the primary factor affecting diabetes risk
2	BMI	0.162	Higher weight/obesity increases disease likelihood
3	C-peptide	0.13	Reflects beta-cell insulin-secreting capacity, useful for distinguishing T1DM/T2DM
4	Insulin	0.125	A crucial metric for assessing secretion vs. resistance
5	Age	0.09	Older age correlates with increased risk of complications

- Top Variable Rankings
  Visualizing high-contributing features via SHAP (SHapley Additive exPlanations) revealed that glucose, BMI, blood pressure, age, C-peptide, insulin, and family history of diabetes were top drivers of model predictions.
- Interpretation in Image Analysis (CNN)
  For CNN models using retinal images, specific lesion areas (microbleeds, changes around the macula, etc.) were found to significantly influence predictions according to SHAP Summary Plots. This provides valuable insights for early detection of diabetic retinopathy.

2) Model Transparency and Credibility Assessment:

- Feedback from Medical Professionals
  SHAP-based visualizations (Force Plots, Dependence Plots, etc.) increased understanding of the decision rationale, moving away from "black box models" by quantitatively and intuitively displaying each feature's impact on predictions.
- Patient Persuasion and Communication
  Offering detailed causal explanations—for instance, "This patient is categorized as high risk due to very low C-peptide levels and large fluctuations in blood glucose"—received positive feedback for enhancing patient compliance with treatment.

- *ANN (Artificial Neural Network)*: Uses a multi-layer perceptron (MLP) structure, adjusting the number of hidden layers and neurons to learn nonlinear relationships.

- *CNN (Convolutional Neural Network):* Particularly advantageous for analyzing retinal or medical images, automatically extracting feature maps through convolution and pooling operations.

- *LSTM (Long Short-Term Memory)*: A variant of recurrent neural networks (RNN) specialized for time-series data, effectively handling long-term dependencies in blood glucose variability or wearable sensor data.

As models become more complex (ensemble, deep-learning combinations), computing SHAP values becomes more time-consuming. Additionally, it remains challenging to fully capture the intricate interactions among variables.

D. Application Scenarios in Real Healthcare Settings

1) Pilot Testing:

- Summary of Results
  Improved Recall: There was a significant rise in the proportion of confirmed diabetic patients among those predicted at high risk. This was recognized for its medical value in preventing missed diagnoses of patients who could be endangered without additional blood or retinal examinations. Maintained Specificity: Though specificity remained around a medium level, experts noted that in a disease where early prevention is crucial, over-predicting positive cases is "clinically acceptable."
- Further Improvement Requests
  Acquire additional variables reflecting detailed diet and exercise patterns (e.g., meal photo recognition, more precise sensor data on physical activity). Enhance mobile functionality for patients (self-monitoring, medication reminders, diet management, etc.). Optimize the frequency of model updates and strengthen data security (personal information protection, server access control, etc.).

E. Comprehensive Discussion

Taken together, the results show that an ensemble approach combining machine learning and deep learning methods is highly effective for improving diabetes prediction accuracy. A significant achievement was boosting Recall by supplementing the scarce positive data via GAN-based augmentation. Moreover, by integrating insulin and C-peptide information, the model more accurately distinguished Type 1 vs. Type 2 diabetes and better detected advanced Type 2 cases in a timely manner.

Above all, applying explainable AI (XAI) techniques such as SHAP played a pivotal role in enhancing credibility and utility in clinical practice. Providing concrete evidence of "why a certain patient is classified as high risk" led to more transparent, persuasive decision-making processes for both healthcare professionals and patients.

Future research should include additional clinical trials, long-term follow-up studies, and expansion to diverse patient groups (varying ages, comorbidities, etc.) to further validate the model's safety and efficacy. Moreover, extending the application to real-time analysis in big data environments and developing interactive services (apps, chatbots, etc.) for patient-clinician communication may significantly contribute to personalized diabetes management and preventive medicine.

F. Interpretation and Significance of the Research Findings

1) Significance of Improved Performance in Early Diabetes Prediction Models:

In this study, the proposed ensemble models—based on machine learning and deep learning—showed overall improvements in major performance indicators (Accuracy, Precision, Recall, etc.) compared to existing models. Notably, the ensemble techniques (Bagging, Boosting, Stacking) that combine multiple models compensated for the limitations of individual algorithms and maximized predictive accuracy by integrating various forms of medical data (numerical, image, time-series). Given that medical data often exhibit heterogeneous and complex characteristics, this improvement carries substantial practical significance.

Moreover, this research introduced insulin and C-peptide data to incorporate the pathological differences between Type 1 and Type 2 diabetes more precisely at the model level. This approach helps overcome the limitations of previous models that primarily focused on Type 2 diabetes. In the long run, it is expected to contribute to the development of more advanced models in precision medicine by comprehensively accounting for beta-cell dysfunction, autoimmunity, and insulin resistance.

2) Clinical Applicability in the Healthcare Field:

This study demonstrated practical clinical applicability by employing data augmentation (via GAN) and XAI techniques (via SHAP). By leveraging data augmentation to mitigate class imbalance in cases of rare diseases or specific patient groups (e.g., positive diabetes cases), the model can more accurately identify borderline or minority instances that would otherwise be easily overlooked. This effect is anticipated to be particularly beneficial in situations where the number of patients with Type 1 diabetes is relatively small.

Additionally, by introducing XAI techniques to make the prediction process interpretable, both clinicians and patients can reduce their distrust in "black box" models and understand the rationale behind decisions. Feedback from pilot testing in real hospital settings indicated that SHAP-based visualizations provided healthcare professionals with patient-specific risk-factor analyses, which aided personalized treatment and patient persuasion.

Furthermore, a trial implementation of a real-time prediction system (web-based dashboard) demonstrated feasible levels of efficiency regarding response time and data throughput, suggesting that personalized alerts and preventive management are practically achievable.

G. Limitations of the Findings and Research Constraints

1) Sample Size and Data Bias:

Although this study utilized various data sources—Pima Indian dataset, domestic hospital EHR, and wearable device data—it remains difficult to completely rule out potential biases toward specific population groups. The Pima Indian dataset, which is centered on the U.S. population, may have genetic and lifestyle differences that do not directly translate to Asian populations. Moreover, the level of data standardization varies among Korean hospitals, posing a risk that different preprocessing methods across samples could lead to performance discrepancies. Hence, there is a need for broader population coverage, standardized protocols for data collection, and consistent preprocessing procedures to ensure more robust and generalizable results.

2) Limitations in Model Interpretability (Scope of XAI Techniques):

Despite incorporating explainable AI (XAI) methods like SHAP, fully interpreting complex ensemble or multimodal deep learning models remains challenging. When strong interactions exist among variables (e.g., C-peptide × insulin × specific genetic variants), SHAP alone may not clearly capture or explain all these interactions. Additionally, when time-series data (LSTM) is combined with image data (CNN), using only partial XAI approaches can make it difficult to intuitively depict the entire prediction process. Therefore, additional visualization strategies, medical education materials, and UI/UX enhancements are required to help patients and non-specialists adequately understand these interpretive results.

3) Expansion Constraints Due to Differences in Domestic and International Environments:

If the proposed model is optimized for a specific country or healthcare system (e.g., Korea or the U.S.), performance may degrade when extended to other regions or nations. For instance, Koreans often have a higher carbohydrate intake and lower obesity indices than Western populations, yet diabetes rates are rising rapidly. A model trained on predominantly Western data may not adequately reflect such nuances. This indicates a limitation—and a future challenge—that multinational, multi-institutional research collaborations are needed to design region-specific models.

H. Future Research Directions

1) Expansion to Predicting Other Chronic Diseases:

The ensemble methods, GAN-based augmentation, and XAI approaches presented in this study can also be extended to early prediction of hypertension, cardiovascular diseases, chronic kidney diseases, cancer, and other chronic illnesses. Research that focuses on multimorbid patients with multiple chronic conditions and holistically analyzes the correlations among different risk factors to develop multi-disease prediction models would be highly valuable both academically and industrially.

2) Use of Additional Data (Genetic Information, Physiological Signals, etc.):

Particularly in diseases with strong autoimmune or genetic components (e.g., Type 1 diabetes), integrating genetic mutation data can significantly enhance the accuracy of risk prediction. While this study primarily examined heart rate and physical activity, future work could integrate various signals such as blood pressure, oxygen saturation, sleep patterns, and stress indices to develop dynamic predictive models. Linking this with continuous glucose monitoring (CGM) technology would be especially beneficial, allowing real-time monitoring of blood glucose fluctuations and continuous updates of risk assessments.

3) Model Lightweighting and Integration with Mobile/Wearable Devices:

Research is needed on techniques like knowledge distillation or model compression so that the model can be reduced in size for real-time inference on hospital servers or mobile devices. By having patients directly enter self-measured blood glucose levels, dietary habits, and exercise data into an app, the model could immediately compute risk levels and alert healthcare providers as necessary. With periodic data collection from smartwatches, CGM (Continuous Glucose Monitoring) devices, etc., the model could learn or make inferences on a streaming basis, offering early-warning notifications (e.g., "Risk of Rapid Blood Glucose Spike") to patients.

I. Overall Conclusion and Implications

This study utilized data augmentation (GAN) and XAI methods (SHAP, etc.) in conjunction with machine learning and deep learning ensembles, thereby enhancing both the performance and interpretability of early diabetes prediction models. As a result, the model effectively addresses both Type 1 and Type 2 diabetes, and pilot tests in clinical settings empirically confirmed its potential to support medical decision-making.

However, questions remain regarding data discrepancies across regions and institutions, interpretability of complex models, and global applicability. Therefore, future steps should include the collection of large-scale standardized datasets, international collaborative research, and more refined XAI techniques. Moreover, integrating genetic information and a wide range of wearable/physiological signals to build a personalized management platform could further advance early diagnosis and preventive systems—not just for diabetes but for other chronic diseases as well.

IV. Conclusion

The primary goal of this study was to develop an early diabetes prediction model using a variety of machine learning and deep learning techniques, and to examine its feasibility in real-world healthcare environments. To this end, we integrated and preprocessed the Pima Indian dataset, EHR data from domestic hospitals, and wearable device data. We compared and evaluated a wide range of algorithms, including SVM, KNN, Random Forest, XGBoost, ANN, CNN, and LSTM, followed by hyperparameter optimization and the introduction of ensemble methods. We also employed GAN to address class imbalance problems and incorporated insulin and C-peptide data to reflect the differing pathogenesis of Type 1 and Type 2 diabetes in the model.

The study outcomes are summarized by employing multi-ensemble models, Accuracy, Recall, and Precision improved by roughly 2–5 percentage points or more compared to single models. The addition of GAN-based synthetic data enhanced the model's ability to identify the positive class (patients with diabetes). Using SHAP, we visualized the model's prediction processes and provided explanation tools that medical professionals can easily understand.

Hence, this research confirmed the feasibility and effectiveness of applying cutting-edge AI methods in the early prediction of diabetes, and we consider the initial objectives ("increased accuracy" and "ensuring explainability") to have been meaningfully achieved. Prediction accuracy surpassed 90% in some ensemble models. In precision and recall, each reached around 90%, improving by 5–10 percentage points compared to previous results. In F1-score, this study maintained over 0.90, indicating balanced model performance. In effect of data augmentation, when using GAN-generated synthetic data, recall for the positive class rose by up to 6 percentage points.

By employing SHAP-based visualization tools, major variables and underlying prediction evidence were made available in an easily understandable format for healthcare professionals and patients. We built a web-based dashboard and alert system, gathering positive feedback from pilot tests with clinical staff. After incorporating feedback from endocrinologists, a real-time prediction system was prototyped, demonstrating both sub-second response times and satisfactory accuracy.

B. Implication and Limitations

This study, which integrates pharmacology, medicine, and engineering, serves as a reference for research into predictive models not only for diabetes but also for various chronic diseases such as hypertension, heart disease, and renal disease. Subsequent research can explore the application of other state-of-the-art machine learning and deep learning techniques, building on the datasets and algorithm evaluation framework established here. By incorporating detailed biological information such as insulin and C-peptide, we have established a basis for providing diagnostic and treatment plans optimized for each patient's unique characteristics.

Hospitals and healthcare companies can develop diabetes risk screening systems, clinical decision-support tools, and patient self-management apps based on the models proposed in this study. Health insurance companies and pharmaceutical firms can leverage risk assessment results to design customized insurance products or use them as additional indicators in targeting new drug development. By integrating data in real time from smartwatches or continuous glucose monitoring (CGM) devices, these models could evolve into diverse service offerings for personalized health management.

Public health authorities could reference the study's models to identify high-risk groups and develop phase-specific intervention strategies. Early diagnoses that lower the incidence of complications can, over the long term, reduce overall national healthcare expenditures and mitigate the decline in patients' quality of life. As further research outcomes accumulate, there is a greater likelihood that national-level initiatives in big data standardization, legislation, and policy improvements will emerge.

Diabetes is a representative chronic disease that exhibits wide variability in genetic predispositions, lifestyle factors, age groups, and the risk of complications among individuals. Rather than classifying and managing patients based on a single criterion, it is more effective to use patient-tailored medical AI models that propose optimized treatments and preventive strategies for each individual. The integrated, ensemble-based approach presented in this study can serve as a significant springboard for developing such personalized models.

Building on this study's results, further work on predicting drug responses, monitoring side effects, and investigating combination drug therapies could help optimize prescriptions and improve patient adherence. With greater model interpretability, clinicians can reference the predictions in their practice, provide detailed explanations to patients and caregivers, and make faster, more accurate decisions.

By acquiring patient data from various races and cultures, the model can be refined for broader generalization and international dissemination. While utmost caution is required to protect personal information, establishing a secure data-sharing system for research purposes could enable large-scale validation and the development of innovative algorithms. This not only fosters academic and industrial progress but also extends treatment opportunities to patients worldwide, thus serving the public good.

Acknowledgment

This paper was supported by Sunchon National University Research Fund in 2024.(Grant number: 2024-0436)

References

[1] I. A. Islam and M. I. Milon, "Diabetes Prediction: A Deep Learning Approach," *International Journal of Information Engineering and Electronic Business*, vol. 11, pp. 21–27, 2019, doi:10.5815/ijieeb.2019.02.03.

[2] J. P. Kandhasamy and S. Balamurali, "Performance Analysis of Classifier Models to Predict Diabetes Mellitus," Procedia Computer Science, vol. 47, pp. 45–51, 2019, doi: 10.1016/j.procs.2015.03.182.

[3] M. Radja and A. W. R. Emanuel, "Performance Evaluation of Supervised Machine Learning Algorithms Using Different Data Set Sizes for Diabetes Prediction," in Proc. 5th Int. Conf. Science in Information Technology (ICSITech), 2019, doi:10.1109/icsitech46713.2019.8987479.

[4] B. G. Choi, S. W. Rha, S. W. Kim, J. H. Kang, J. Y. Park, Y. K. Noh, and S. I. Choi, "Machine Learning for the Prediction of New-Onset Diabetes Mellitus During 5-Year Follow-up in Nondiabetic Patients with Cardiovascular Risks," Yonsei Medical Journal, vol. 60, no. 2, pp. 191–199, 2019, doi: 10.3349/ymj.2019.60.2.191.

[5] R. Akula, N. Nguyen, and I. Garibay, "Supervised Machine Learning-Based Ensemble Model for Accurate Prediction of Type 2 Diabetes," in Proc. SoutheastCon, 2019, doi:10.1109/southeastcon42311.2019.9020358.

[6] Z. Xie, O. Nikolayeva, J. Luo, and D. Li, "Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques," Preventing Chronic Disease, vol. 16, p. E130, 2019, doi:10.5888/pcd16.190109.

[7] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive Models for Diabetes Mellitus Using Machine Learning Techniques," BMC Endocrine Disorders, vol. 19, no. 1, p. 101, 2019, doi: 10.1186/s12902-019-0436-6.

[8] H. Abbas, L. Alic, M. Erraguntla, J. Ji, M. AbdulGhani, Q. Abbasi, and M. Qaraqe, "Predicting Long-Term Type 2 Diabetes with Support Vector Machine Using Oral Glucose Tolerance Test," PLoS ONE, vol. 14, no. 7, p. e0219636, 2019, doi: 10.1371/journal.pone.0219636.

[9] B. Farran, R. Al-Wotayan, H. Alkandari, D. Al-Abdulrazzaq, A. Channanath, and T. A. Thanaraj, "Use of Non-Invasive Parameters and Machine-Learning Algorithms for Predicting Future Risk of Type 2 Diabetes: A Retrospective Cohort Study of Health Data from Kuwait," Frontiers in Endocrinology, vol. 10, p. 624, 2019, doi:10.3389/fendo.2019.00624.

[10] X. L. Xiong, R. X. Zhang, Y. Bi, W. H. Zhou, Y. Yu, and D. L. Zhu, "Machine Learning Models in Type 2 Diabetes Risk Prediction: Results from a Cross-Sectional Retrospective Study in Chinese Adults," Current Medical Science, vol. 39, no. 4, pp. 582–588, 2019, doi:10.1007/s11596-019-2077-4.

[11] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A Data-Driven Approach to Predicting Diabetes and Cardiovascular Disease with Machine Learning," BMC Medical Informatics and Decision Making, vol. 19, no. 1, p. 211, 2019, doi: 10.1186/s12911-019-0918-5.

[12] Y. Liu, S. Ye, X. Xiao, C. Sun, G. Wang, G. Wang, and B. Zhang, "Machine Learning for Tuning, Selection, and Ensemble of Multiple Risk Scores for Predicting Type 2 Diabetes," Risk Management and Healthcare Policy, vol. 12, pp. 189–198, 2019, doi:10.2147/rmhp.s225762.

[13] K. Leerojanaprapa and K. Sirikasemsuk, "Comparison of Bayesian Networks for Diabetes Prediction," in International Conference on Computer, Communication and Computational Sciences (IC4S), Bangkok, Thailand, 2018, Advances in Intelligent Systems and Computing, vol. 924, pp. 425–434, doi: 10.1007/978-981-13-6861-5_37.

[14] N. Sneha and T. Gangil, "Analysis of Diabetes Mellitus for Early Prediction Using Optimal Features Selection," Journal of Big Data, vol. 6, no. 1, p. 13, 2019, doi: 10.1186/s40537-019-0175-6.

[15] 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, doi:10.1007/s40200-020-00520-5.

[16] H. Zhou, R. Myrzashova, and R. Zheng, "Diabetes Prediction Model Based on an Enhanced Deep Neural Network," EURASIP Journal on Wireless Communications and Networking, vol. 2020, no. 1, p. 148, 2020, doi: 10.1186/s13638-020-01765-7.

[17] M. Seera and C. P. Lim, "A Hybrid Intelligent System for Medical Data Classification," Expert Systems with Applications, vol. 41, no. 5, pp. 2239–2249, 2020, doi: 10.1016/j.eswa.2013.09.022.

[18] I. Sarker, M. Faruque, H. Alqahtani, and A. Kalim, "k-Nearest Neighbor Learning Based Diabetes Mellitus Prediction and Analysis for e-Healthcare Services," EAI Endorsed Transactions on Scalable Information Systems, 2020, doi: 10.4108/eai.13-7-2018.162737.

[19] A. Cahn, A. Shoshan, T. Sagiv, R. Yesharim, R. Goshen, V. Shalev, and I. Raz, "Prediction of Progression from Prediabetes to Diabetes: Development and Validation of a Machine Learning Model," Diabetes/Metabolism Research and Reviews, vol. 36, no. 2, p. e3252, 2020, doi: 10.1002/dmrr.3252.

[20] R. García-Carretero, L. Vigil-Medina, I. Mora-Jiménez, C. Soguero-Ruiz, O. Barquero-Pérez, and J. Ramos-López, "Use of a k-Nearest Neighbors Model to Predict the Development of Type 2 Diabetes Within 2 Years in an Obese, Hypertensive Population," Medical & Biological Engineering & Computing, vol. 58, no. 5, pp. 991–1002, 2020, doi: 10.1007/s11517-020-02132-w.

[21] L. Zhang, Y. Wang, M. Niu, C. Wang, and Z. Wang, "Machine Learning for Characterizing Risk of Type 2 Diabetes Mellitus in a Rural Chinese Population: The Henan Rural Cohort Study," Scientific Reports, vol. 10, p. 4406, 2020, doi: 10.1038/s41598-020-61123-x.

[22] A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Ali, "Intelligent Machine Learning Approach for Effective Recognition of Diabetes in e-Healthcare Using Clinical Data," Sensors, vol. 20, no. 9, p. 2649, 2020, doi: 10.3390/s20092649.

[23] T. Yang, L. Zhang, L. Yi, H. Feng, S. Li, H. Chen, J. Zhu, J. Zhao, Y. Zeng, H. Liu, et al., "Ensemble Learning Models Based on Noninvasive Features for Type 2 Diabetes Screening: Model Development and Validation," JMIR Medical Informatics, vol. 8, no. 6, p. e15431, 2020, doi: 10.2196/15431.

[24] H. S. Ahn, J. H. Kim, H. Jeong, J. Yu, J. Yeom, S. H. Song, S. S. Kim, I. J. Kim, and K. Kim, "Differential Urinary Proteome Analysis for Predicting Prognosis in Type 2 Diabetes Patients With and Without Renal Dysfunction," International Journal of Molecular Sciences, vol. 21, no. 12, p. 4236, 2020, doi: 10.3390/ijms21124236.

[25] Y. Tang, R. Gao, H. H. Lee, Q. S. Wells, A. Spann, J. G. Terry, J. J. Carr, Y. Huo, S. Bao, B. A. Landman, et al., "Prediction of Type II Diabetes Onset With Computed Tomography and Electronic Medical Records," in Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures, Springer, 2020, pp. 13–23, doi:10.1007/978-3-030-60946-7_2.

[26] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and Prediction of Diabetes Disease Using Machine Learning Paradigm," Health Information Science and Systems, vol. 8, no. 1, p. 7, 2020, doi: 10.1007/s13755-019-0095-z.

[27] S. Jain, "A Supervised Model for Diabetes Divination," Biosc Biotech Res Comm, vol. 13, no. 14, pp. 315–318, 2020, doi:10.21786/bbrc/13.14/7.

[28] P. B. K. Chowdary and R. U. Kumar, "An Effective Approach for Detecting Diabetes Using Deep Learning Techniques Based on Convolutional LSTM Networks," International Journal of Advanced Computer Science and Applications, vol. 12, no. 8, pp. 519–525, 2021, doi: 10.14569/ijacsa.2021.0120466.

[29] J. A. Mat Jizat, R. A. Rahim, S. Harun, M. S. Khan, M. S. M. Rizam, and N. M. Saad, "Evaluation of the Machine Learning Classifier in Wafer Defects Classification," ICT Express, vol. 7, no. 4, pp. 535–539, 2021, doi: 10.1016/j.icte.2021.04.007.

[30] J. Liu, L. Fan, Q. Jia, L. Wen, and C. Shi, "Early Diabetes Prediction Based on Stacking Ensemble Learning Model," in Proc. 33rd Chinese Control and Decision Conference (CCDC), 2021, pp. 5722–5727, doi:10.1109/CCDC52312.2021.9601932.

[31] L. Fregoso-Aparicio, J. Noguez, L. Montesinos, and J. A. García-García, "Machine Learning and Deep Learning Predictive Models for Type 2 Diabetes: A Review," Diabetology & Metabolic Syndrome, vol. 13, p. 767, 2021, doi: 10.1186/s13098-021-00767-9.

[32] A. C. Lyngdoh, N. A. Choudhury, and S. Moulik, "Diabetes Disease Prediction Using Machine Learning Algorithm," in Proc. IEEE-EMBS Conf. Biomedical Engineering and Sciences (IECBES), 2021, pp. 935–940, doi: 10.1109/IECBES48179.2021.9398759.

[33] A. Tack, B. Preim, and S. Zachow, "Fully Automated Assessment of Knee Alignment from Full-Leg X-Rays Employing a "YOLOv4 and ResNet Landmark Regression Algorithm" (YARLA): Data from the Osteoarthritis Initiative," Computer Methods and Programs in Biomedicine, vol. 203, p. 106080, 2021, doi:10.1016/j.cmpb.2021.106080.

[34] J. J. Boutilier, T. C. Y. Chan, M. Ranjan, and S. Deo, "Risk Stratification for Early Detection of Diabetes and Hypertension in Resource-Limited Settings: Machine Learning Analysis," Journal of Medical Internet Research, vol. 23, no. 1, p. e20123, 2021, doi:10.2196/20123.

[35] J. Li, Q. Chen, X. Hu, P. Yuan, L. Cui, T. Jiang, and J. Ma, "Establishment of Noninvasive Diabetes Risk Prediction Model Based on Tongue Features and Machine Learning Techniques," International Journal of Medical Informatics, vol. 149, p. 104429, 2021, doi:10.1016/j.ijmedinf.2021.104429.

[36] H. B. Kibria, M. Nahiduzzaman, M. O. F. Goni, M. Ahsan, and J. Haider, "An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI," Sensors, vol. 22, no. 19, p. 7268, 2022, doi: 10.3390/s22197268.

[37] A. Dutta, M. K. Hasan, M. Ahmad, M. A. Awal, M. A. Islam, M. Masud, and H. Meshref, "Early Prediction of Diabetes Using an Ensemble of Machine Learning Models," International Journal of Environmental Research and Public Health, vol. 19, no. 19, p. 12378, 2022, doi: 10.3390/ijerph191912378.

[38] H. Wei, J. Sun, W. Shan, W. Xiao, B. Wang, M. Hu, X. Wang, and Y. Xia, "Environmental Chemical Exposure Dynamics and Machine Learning Based Prediction of Diabetes Mellitus," Science of the Total Environment, vol. 806, p. 150674, 2022, doi:10.1016/j.scitotenv.2021.150674.

[39] S. M. Ganie, P. K. D. Pramanik, M. B. Malik, S. Mallik, and H. Qin, "An Ensemble Learning Approach for Diabetes Prediction Using Boosting Techniques," Frontiers in Genetics, 2023, doi:10.3389/fgene.2023.1252159.

[40] M. F. Aslan and K. Sabanci, "A Novel Proposal for Deep Learning-Based Diabetes Prediction," Diagnostics, vol. 13, no. 4, p. 796, 2023, doi: 10.3390/diagnostics13040796.

[41] K. Abnoosian, R. Farnoosh, and M. H. Behzadi, "Prediction of Diabetes Disease Using an Ensemble of Machine Learning Multi Classifier Models," BMC Bioinformatics, vol. 24, p. 5465, 2023, doi:10.1186/s12859-023-05465-z.

[42] A. E.-S. El-Bashbishy and H. M. El-Bakry, "Pediatric Diabetes Prediction Using Deep Learning," Scientific Reports, 2024, doi:10.1038/s41598-024-51438-4.