Submitted: 29 May, 2024 Accepted: 22 August, 2024 Published: 30 November, 2024 DOI:10.22514/jomh.2024.184

JOMH
Journal of Men's Health

ORIGINAL RESEARCH

Explainable stacking ensemble with feature tokenizer transformers for men's diabetes prediction

Vinh Quang Tran[1], Younsung Choi[2,*,†], Haewon Byeon[2,*,†]

[1]Department of Digital Anti-Aging Healthcare (BK21), Inje University, 50834 Gimhae, Republic of Korea
[2]Department of AI-Software, Inje University, 50834 Gimhae, Republic of Korea

***Correspondence**
cys2020@inje.ac.kr
(Younsung Choi);
bhwpuma@naver.com
(Haewon Byeon)

[†] These authors contributed equally.

Abstract

Diabetes is a leading global health concern, with millions of deaths linked to diabetes and related complications according to the World Health Organization (WHO). Early and accurate prediction is crucial for effective management. This study investigates the potential of a stacking ensemble approach for predicting diabetes in men (n = 5598). The ensemble leverages a Feature Tokenizer transformer, a deep learning technique, alongside various machine learning models. SHAP (SHapley Additive exPlanations) is used to enhance model interpretability. Compared to other stacking methods and standalone models, the proposed ensemble with a Random Forest meta-classifier, XGBoost, Feature Tokenizer Transformers (FT-Transformer) and LightGBM achieved superior performance (accuracy: 0.8786, precision: 0.7989, recall: 0.8171, F1-score: 0.8079, Area Under the Curve (AUC): 0.8618). These findings suggest that stacking ensembles with deep learning and explainable artificial intelligent (AI) hold promise for improving diabetes prediction in men, potentially leading to better clinical decision-making and patient outcomes.

Keywords

Feature tokenizer; Men's health; Diabetes; Explainable artificial intelligent

1. Introduction

Diabetes mellitus (DM), commonly referred to as diabetes, encompasses a spectrum of metabolic disorders characterized by chronic hyperglycemia. These disorders typically arise from disruptions in either insulin secretion, insulin action, or a combination of both factors [1]. The World Health Organization (WHO) has documented a significant increase in diabetes-related mortality, with a 70% rise since the year 2000. This alarming trend has elevated diabetes to one of the ten leading causes of death globally [2]. In 2019 alone, an estimated two million deaths were attributed to diabetes and its associated kidney complications [3]. Diabetes manifests in three primary forms: Type 1 (T1DM), Type 2 (T2DM) and Gestational Diabetes (GD). While all three forms elevate blood sugar levels, their underlying etiologies differ. T1DM, an autoimmune disease, results from the body's attack on insulin-producing cells. Typically diagnosed in childhood, T1DM necessitates lifelong insulin supplementation for management [4]. In contrast, T2DM arises from either insulin resistance or insufficient insulin production. While the prevalence of T2DM is on the rise for both sexes, men tend to be diagnosed at a younger age and with a lower body fat mass compared to women. Notably, research indicates an estimated 17.7 million greater prevalence of diabetes mellitus in men compared to women [5].

While ongoing research actively explores advancements in diabetes treatment and potential cures, a universally accepted cure remains elusive within the scientific community. Fortunately, effective management strategies exist, often incorporating diet, exercise and sometimes medication. Within the realm of preventative and control measures, as advocated by scientists, physical activity plays a primary and pivotal role alongside medicinal interventions and dietary modifications [6]. However, early detection and diagnosis of diabetes remain crucial in contemporary healthcare settings. Timely diagnosis not only aids in preventing complications but also significantly reduces the risk of developing other chronic conditions like kidney disease, heart attack and stroke. Given the complexities of diabetes and the critical need for personalized care plans, prioritizing resources for disease prediction in large populations is essential. This is particularly important considering the phenomenon of physician burnout, often attributed to the demands of electronic health record (EHR) systems [7]. Herein lies the significance of researcher-developed tools and systems designed to assist in the prediction and classification of diabetes.

Diabetes prediction has been a well-studied area in healthcare, encompassing diagnosis, classification and treatment strategies. Recent research leverages various machine learning (ML) and deep learning (DL) algorithms to identify and predict diabetes [8–10]. These algorithms have yielded significant improvements over traditional and basic ML methods. However, while DL excels in handling diverse data types like natural language, audio and images, its application to tabular datasets presents challenges. These challenges include difficulties in optimizing DL models due to missing data, the presence of mixed feature types (numerical, ordinal, categorical), and the lack of inherent structural knowledge compared to well-defined structures in text or image data [11]. Nevertheless, Shwartz-Ziv *et al.* [12] (2022) highlight those ensembles combining XGBoost models with DL models often outperform standalone XGBoost models across various datasets.

While existing research often focuses on both genders, women or gestational diabetes alone [13–15], a gap exists in developing models specifically for male diabetes classification. To address this gap, we propose an ensemble model tailored for tabular data. This ensemble combines a Feature Tokenizer Transformer (FTT) [16], featuring a Feature Tokenizer component and adapted Transformer layers [17], with XGBoost [18], LightGBM [19] and Random Forest (RF) [20]. For comparative purposes, we additionally employed other ML and DL methods like TabNet [21] and CatBoost [22] on a large dataset from the Korea National Health and Nutrition Examination Survey (KNHANES) for male diabetes prediction. Furthermore, to create a robust classification model with Explainable Artificial Intelligence (XAI), we utilized SHapley Additive exPlanations (SHAP) [23]. This approach, recognized in numerous diabetes classification studies [8, 24–28], enhances model interpretability, revealing the inner workings of both DL and ML models. This is particularly valuable for our proposed method, which might be initially complex to understand. By integrating SHAP, the model can provide insights not only for data scientists but also for physicians, potentially guiding future research directions and clinical applications.

This paper is organized as follows. The next section (Section 2) reviews relevant background literature. Section 3 details the proposed approach, including a description of the datasets and algorithms employed. Section 4 presents the model's performance and offers a concise rationale for the chosen methods. Section 5 delves into interpretability using XAI-SHAP techniques. Finally, Section 6 concludes by summarizing the key findings and outlining promising directions for future research.

2. Literature review

A significant surge in research has emerged in recent years focusing on the application of ML and DL models for diabetes patient identification. Ensemble techniques, which combine multiple individual models to enhance prediction performance, have gained particular traction in this domain. Notably, Dutta *et al.* [29] (2022) achieved a high accuracy of 0.735 and an area under the Receiver Operating Characteristic (ROC) curve (AUC) of 0.832 using an ensemble pipeline. This pipeline incorporated a weighted ensemble of RF, Naive Bayes (NB), XGBoost, Decision Tree (DT), and LightGBM (LGB) with grid search optimization, missing value imputation, feature selection, and K-fold cross-validation. Furthermore, their model achieved this accuracy using only four to five interpretable features: body mass index (BMI), age, average systolic and diastolic blood pressure and occupation.

While ensemble models offer strong performance, alternative approaches exist. Chang *et al.* [30] (2023) explored using an artificial neural network (ANN) on the 7th KNHANES dataset (2016–2018), focusing on 11 nutritional intake features and achieving an accuracy of 0.813. Similarly, Choi *et al.* [31] (2023) employed a support vector machine (SVM) model on the 5th KNHANES data, attaining an AUC of 0.731 for prediabetes prediction. Additionally, Kumari *et al.* [32] (2021) utilized a soft voting ensemble combining RF, Logistic Regression (LR), and NB, achieving an accuracy of 79.04% along with good precision and recall.

Beyond prediction accuracy, XAI methods like SHAP, Local Interpretable Model-Agnostic Explanations (LIME) and Explain Like I'm 5 (ELI5) have been increasingly employed to provide insights into the decision-making processes of ML and DL models. These methods can be further enhanced by user interface integration, as shown in Table 1.

3. Materials and methods

In the pursuit of constructing a robust and interpretable stacking ensemble model for the prediction of diabetes, this study undertook a series of systematic steps, outlined in Fig. 1.

3.1 Data-set description

This study leveraged data from the Korea National Health and Nutrition Examination Survey (KNHANES) conducted between 2019 and 2022 (further details on KNHANES design and data can be found in [37]). The KNHANES, a cross-sectional survey by the Korea Centers for Disease Control and Prevention (KCDC), involved 13,152 men. It gathers information on various aspects of health through three components: health interviews, health examinations, and a dietary survey. These components provide data on socioeconomic status, health behaviors, quality of life, healthcare use, anthropometric measurements, biochemical and clinical profiles for non-communicable diseases and dietary intake.

3.2 Data preprocessing

3.2.1 Data cleaning and splitting for model training

Data preprocessing is crucial for building robust models, especially when dealing with tabular data [38]. In this study, data from four years (presumably from KNHANES) was combined with information on 13,152 men, resulting in a dataset with 632 features. From these features, the column labeled "HE_DM_HBA1C" was selected as the target variable. This column indicates whether a participant has diabetes based on three criteria: fasting blood glucose above 126 mg/dL, physician diagnosis, or use of hypoglycemic agents (medications that lower blood sugar).

Since the data originated from a survey, some questions were optional, resulting in uneven response rates across certain columns. Consequently, these columns had missing data exceeding a 50% threshold. Fortunately, they contained non-essential or supplementary information, justifying their re-

TABLE 1. Recent research on XAI for DM diagnosis.

Article	Dataset	Models	XAI method	F1-Score	Accuracy	Other metrics
Guha *et al.* [24] (2020)	ESDRPD dataset (520 patients)	Random Forest	SHAP/LIME/ELI5	0.95	95.00%	Precision: 0.95 Recall: 0.94 AUC: 0.98
Kibria *et al.* [28] (2022)	PIMA [33]	Soft Voting Ensemble of XGBoost and Random Forest	LIME/SHAP	0.95	90.00%	Precision: 0.88 Recall: 0.89 AUC: 0.95
Joseph *et al.* [8] (2022)	PIMA dataset and ESDRPD dataset	Bayesian-Optimized Hyperparameter TabNet	SHAP/LIME/ TabNet/ELI5	0.88	92.20%	Precision: 0.86 Specificity: 0.95
Vishwarupe *et al.* [26] (2022)	Local Pune dataset (1367 patients)	Random Forest	SHAP/LIME/ELI5	N/A	82.23%	N/A
Curia *et al.* [34] (2023)	Dhaka dataset (306 patients)	XGBoost	LIME	1.00	100.00%	Precision: 1.00
Tasin *et al.* [27] (2023)	PIMA	XGBoost with ADASYN	LIME/SHAP	0.81	88.50%	Precision: 0.82 Recall: 0.80
Dharmarathne *et al.* [35] (2024)	Public diabetes dataset [36]	Self-explainable interface with XGBoost	SHAP	0.65	77.00%	Precision: 0.60 Recall: 0.73 AUC: 0.82

XGBoost: XGBoost Classifier; Random Forest: Random Forest Classifier; XGBoost with ADASYN: XGBoost with ADASYN over-sampling method. XAI: Explainable Artificial Intelligence; ESDRPD: Early-stage diabetes risk prediction dataset; SHAP: SHapley Additive exPlanations; LIME: Local Interpretable Model-Agnostic Explanations; ELI5: Explain Like I'm 5; AUC: Area Under the Curve; PIMA: Pima Indian Diabetes Dataset; N/A: Not Available; ADASYN: Adaptive Synthetic Sampling.

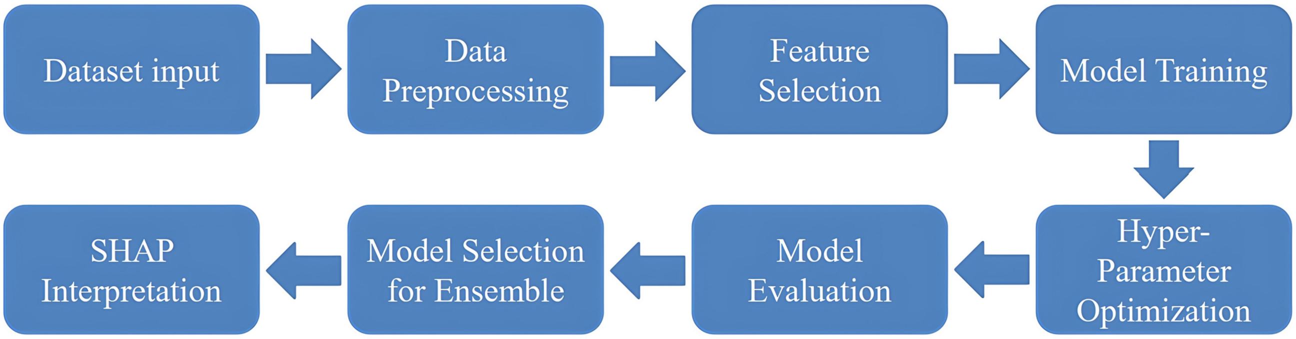

FIGURE 1. Schematic representation of the analysis process in this study. The data was first loaded and then prepared (preprocessed) to ensure its quality and suitability for analysis. After preprocessing, the most important features influencing the target variable were identified through a feature selection process. Next, DL and ML models were built and trained on the preprocessed data with the selected features. Once trained, the model's hyperparameters were adjusted to optimize its performance. Following this optimization, the best performing models were chosen to create an ensemble model. Finally, SHAP interpretation was conducted to gain insights into the ensemble model's predictions. SHAP: SHapley Additive exPlanations.

moval. To address this challenge, any column with missing values exceeding 50% was excluded from the dataset. Furthermore, to ensure data relevance to men's diabetes prediction, we excluded males under 19 years old and those lacking an "HE_DM_HBA1C" value. Additionally, existing diabetes-related columns, such as blood glucose, glycated hemoglobin, and insulin injection data, were excluded. This data cleaning process resulted in a final dataset containing a substantial number of samples (5598) with 195 variables. In the field of ML, cross-validation is a commonly used technique for evaluating model performance on limited and tabular datasets [38]. However, when dealing with DL, particularly for complex models, cross-validation can be computationally expensive and time-consuming [39]. Therefore, we opted for an alternative approach. Following established practices in ML, we employed a stratified 80/10/10 train-test-validation split on the preprocessed data (Géron, 2019) [40]. This approach is particularly recommended for DL models due to their data requirements. The dataset was stratified and split into a training set (80%), a validation set (10%), and a hold-out test set (10%). The training set will be used to build the model, the validation set will be used for hyperparameter tuning, and the hold-out test set will be used for the final evaluation of the model's performance on unseen data.

3.2.2 Feature scaling and missing value handling

To ensure all features contribute equally and improve the model's ability to generalize to unseen data, we addressed the different scales of numerical and categorical variables. We first separated the dataset's features into these two categories. For numerical features, we applied a scaling technique to standardize their values within a common range, like −1 to 1. This prevents features with large scales from dominating the model's predictions and ensures all features have a proportional influence. For example, "age" in years might be scaled to this range. For those with a natural order (like age), we employed encoding by category (*e.g.*, younger, adult, elder) represented by numbers (0, 1, 2). This approach not only improves the interpretability of the model's results but also increases its sensitivity to meaningful variations within these categories, potentially leading to better accuracy and broader applicability. Finally, we addressed columns with missing values exceeding 50%. We will use imputation techniques: mean imputation for numerical data and mode imputation for categorical data.

3.2.3 Data augmentation

Our dataset exhibited class imbalance, a common challenge where one class has significantly more samples than another [41]. In our case, there were 3847 non-diabetic patients compared to 1750 diabetic patients out of 5598 total samples. This imbalance can hinder the performance of standard classifiers. To address this, we employed the Synthetic Minority Oversampling Technique (SMOTE) by Chawla *et al.* [42] (2002). SMOTE tackles class imbalance by creating synthetic samples for the under-represented class (diabetic patients). This is achieved by interpolating between existing minority samples and their nearest neighbors in the feature space. By applying SMOTE to the training dataset, we were able to create a balanced dataset with 2727 samples each for both diabetic and non-diabetic classes. To implement SMOTE, we leveraged the imbalanced-learn library. This library, compatible with scikit-learn, offers a comprehensive suite of tools specifically tailored for addressing classification tasks with imbalanced datasets.

3.2.4 Feature selection

Several ML models possess inherent feature selection mechanisms. Random Forest, XGBoost, LightGBM and CatBoost fall into this category. Conversely, TabNet and FT-Transformer models utilize attention layers to assess feature importance. However, these approaches present limitations. The generated feature importance scores exhibit variability contingent upon the employed classifier. Additionally, solely relying on accuracy for feature inclusion or exclusion might be inadequate [43]. Therefore, a feature selection method that identifies all relevant features, rather than solely focusing on minimal optimal ones, is desirable. Our objective is to achieve a comprehensive understanding of the underlying phenomenon, encompassing all contributing factors, not just non-redundant ones. This approach not only helps us to mitigates the risk of overfitting that can arise from a high number of features.

For this reason, we opted for the BorutaShap method [44] as our preferred approach for significant feature selection within this study. BorutaShap, a Python wrapper method, integrates the Boruta feature selection algorithm [44] with SHAP values. It is specifically designed for seamless operation with tree-based learners such as Random Forest, XGBoost, LightGBM and CatBoost. The Boruta method generates shadow features, which are essentially identical copies of the original features. However, these shadow features incorporate randomization to eliminate any potential associations with the outcome variable. Subsequently, both the original features and their corresponding shadow-shuffled equivalents are incorporated into the tree-based model to forecast the target feature, effectively leveraging the unique capabilities of these specific learners. The algorithm then computes the mean decrease in accuracy (*MD*) for both the shadow-shuffled features and the actual features within (S_{tree}). This calculation is represented by the following formula:

$$MD = \frac{1}{s_{tree}} \textstyle\sum_{s=1}^{s_{tree}} \frac{\sum_{t\in OOB} I(y_t=f(x_t)) - \sum_{t\in OOB} I(y_t=f({x_t}^n))}{|OOBPE|} \quad (1)$$

Within this formula x_t represents predictor variables ($x_t\in R^n$), y_t is the target feature ($y_t \in R$) for n inputs in the set T ($t = 1, 2, ..., T$), the function I is an indicator function, *OOBPE* is the Out-of-Bag Predictive Error, $y_t = f(x_t)$ represents predicted values before permutation, and $y_t = f({x_t}^n)$ represents the predicted values post permutation. By using a two-sided hypothesis test (*t*-test) to compare actual and shadowed values, the algorithm calculates a *Z-score*:

$$Z\text{-}score = \frac{M}{S} \quad (2)$$

Where S denotes the standard deviation of accuracy losses. A specific threshold is established, wherein the *Z-score* of the actual feature must surpass the highest *Z-score* (Z_{max}) obtained from randomized shadow features. Furthermore, to guarantee consistency in the shapely important values (SHAP values), the BorutaShap method compares each original feature to its corresponding shadow feature.

We used the BorutaShap library to find significant features for each model (Random Forest, XGBoost, LightGBM and CatBoost) with default settings and 10-fold cross-validation. BorutaShap was run for 50 iterations. It identified the optimal subset of features by evaluating how much each feature improved the performance of all models compared to using the entire feature set. To assess this improvement, we employed various metrics including accuracy, precision, recall, F1-score and AUC.

3.3 Model architecture

This research utilizes ensemble stacking, a meta-learning technique for classification tasks [45]. In ensemble stacking, multiple base ML models are first trained independently on the entire training dataset. These diverse base learners capture different aspects of the data. Then, the predictions generated by the base models become the new features for a final meta-classifier. This meta-classifier learns to combine the strengths of the base models, resulting in a more robust and accurate prediction model.

In this proposed methodology, we have used the ensemble of ML algorithms such as RF, XGBoost, LightGBM and CatBoost with FT-Transformer. To enhance the efficacy of these models, the Optuna framework [46] was harnessed for hyperparameter fine-tuning, thereby optimizing their predictive performance. The above-mentioned algorithms have been ensembled with a stacking classifier to enhance accuracy. To select meta-model, we make a comprehensive comparison between stacking model for most efficient prediction model for men's diabetes individuals. These algorithms are briefly discussed in this section.

- RF: Random Forest [20], an ensemble method that combines numerous decision trees. Known for its robustness and accuracy, this technique combines the predictions of diverse trees to deliver reliable results, demonstrating resilience against noise and overfitting.

- XGBoost [18] or eXtreme Gradient Boosting, extends gradient boosting by employing a unique regularization term (*e.g.*, L1/L2) and parallel computing to achieve superior accuracy across a diverse range of tasks, including regression, classification and ranking.

- LightGBM [19] builds utpon the Gradient Boosting Decision Tree (GBDT) with innovative techniques like Gradient-based One-Side Sampling and the Histogram-based Algorithm. These methods accelerate training time, reduce memory usage, and ultimately enhance the precision of its GBDT model.

- CatBoost [22] a cutting-edge gradient boosting toolkit, boasts distinct advantages over traditional GBDT models. Unlike traditional gradient boosting models, CatBoost handles categorical features directly, eliminating the need for separate preprocessing steps. This saves time and simplifies the modeling process. Additionally, CatBoost analyzes not only individual features but also their interactions. By considering these feature combinations, the model can capture more complex relationships within the data, potentially leading to more accurate predictions.

- TabNet [21], a robust deep learning model, has demonstrated commendable performance across diverse datasets. The architecture encompasses an encoder module that capitalizes on sequential decision steps to encode features. It employs sparse learned masks to select pertinent features for each row through an attention mechanism. A distinct characteristic of TabNet is its utilization of sparsemax layers, compelling the selection of a compact set of features. This approach deviates from traditional all-or-nothing feature selection, allowing for nuanced decisions via learnable masks. This not only circumvents the rigidity of hard feature thresholds but also enables a soft, differentiable approach to feature selection.

- The FT-Transformer (Feature Tokenizer + Transformer) model [16] adapts the Transformer architecture [17] for tabular data. It simplifies the process by transforming all features (both categorical and numerical) into tokens. These tokens are then fed into a series of stacked Transformer layers, where each layer analyzes the features of individual data points. Within the Transformer component, a special Classification (CLS) token is added and processed alongside the other tokens through multiple layers. This pre-normalization step improves optimization and overall performance. Finally, the model uses the final representation of the (CLS) token for prediction. The FT-Transformer, as described earlier and depicted in Fig. 2.

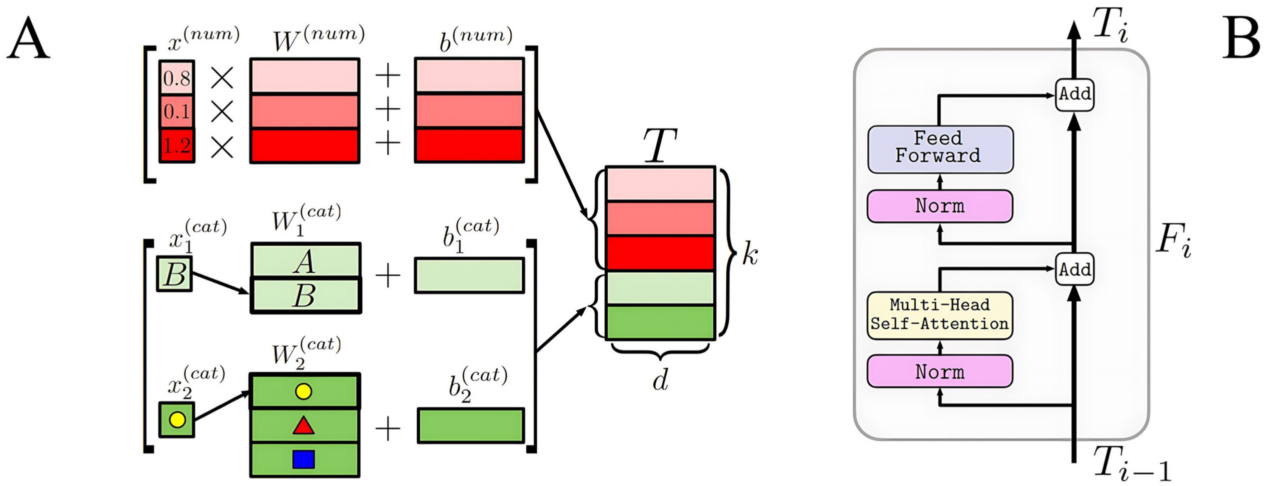

FIGURE 2. Concept of feature tokenizer + transformer illustrated by Gorishniy *et al.* [16]. (A) Feature Tokenizer: This part transforms raw features (in this case, two categorical and three numerical) into tokens for the Transformer to process. (B) Transformer Layer: This layer analyzes the relationships between the tokens, ultimately using them to make predictions.

- Proposed ensemble stacking classifiers method: While traditional ML models excel at finding patterns in data, combining their strengths can lead to even better predictions. Stacking ensemble [45], by combining multiple models, takes a two-stage approach unlike methods like majority voting. In stage one, diverse base models (ML or DL) are trained independently on the entire dataset. These models capture various aspects of the data. The key difference from voting lies in stage two. Stacking uses the predictions from the base models as new features, creating a richer dataset. Each data point now includes the original features plus predictions from each base model. Finally, a new model, called a meta-classifier, is trained on this expanded dataset. This meta-classifier learns from the combined insights of the base models, aiming for superior prediction accuracy compared to individual models. The benefits of stacking ensemble are twofold. First, it leverages diverse models like voting classifiers, but offers a more sophisticated approach. By allowing the meta-classifier to learn from the relationships between original features and base model predictions, stacking can capture more complex data patterns. This can significantly improve prediction performance compared to individual models or simpler ensemble methods. In this research, we implement 10 fold stratified cross-validation to prevent overfitting with optimized parameters for each model. (Fig. 3 shows a visual representation of the proposed methodology). The Python code for implementing the ensemble stacking classifiers method can be found in the **Supplementary material**.

3.4 Mitigating overfitting for robust model performance

This study employed several strategies to prevent overfitting and ensure the generalizability of the proposed ensemble model for diabetes prediction in males. Here, we discuss the specific techniques implemented and their contributions:

- Regularization techniques penalize overly complex models, discouraging them from learning intricate patterns in the training data that might not generalize well to unseen data. In our case, we likely employed L1 or L2 regularization within the LightGBM model (refer to **Supplementary Table 1** for the specific regularization parameter value). This constrains the model's weights and prevents excessive memorization of training data specifics.

- Dropout is a powerful regularization technique specifically designed for neural networks. During training, a random subset of neurons is dropped out with a pre-defined probability. This forces the network to learn robust features that are not overly reliant on any single neuron and encourages the development of diverse internal representations. Our study likely utilized three specific dropout techniques within the FT-Transformer architecture:

◦ Attention dropout: This technique randomly drops out attention weights within the attention layer, promoting a focus on informative features while preventing overfitting to specific attention patterns.

◦ FFN dropout drops out neurons within the feed-forward network (FFN) layers, encouraging the network to learn more robust feature representations and reducing reliance on any

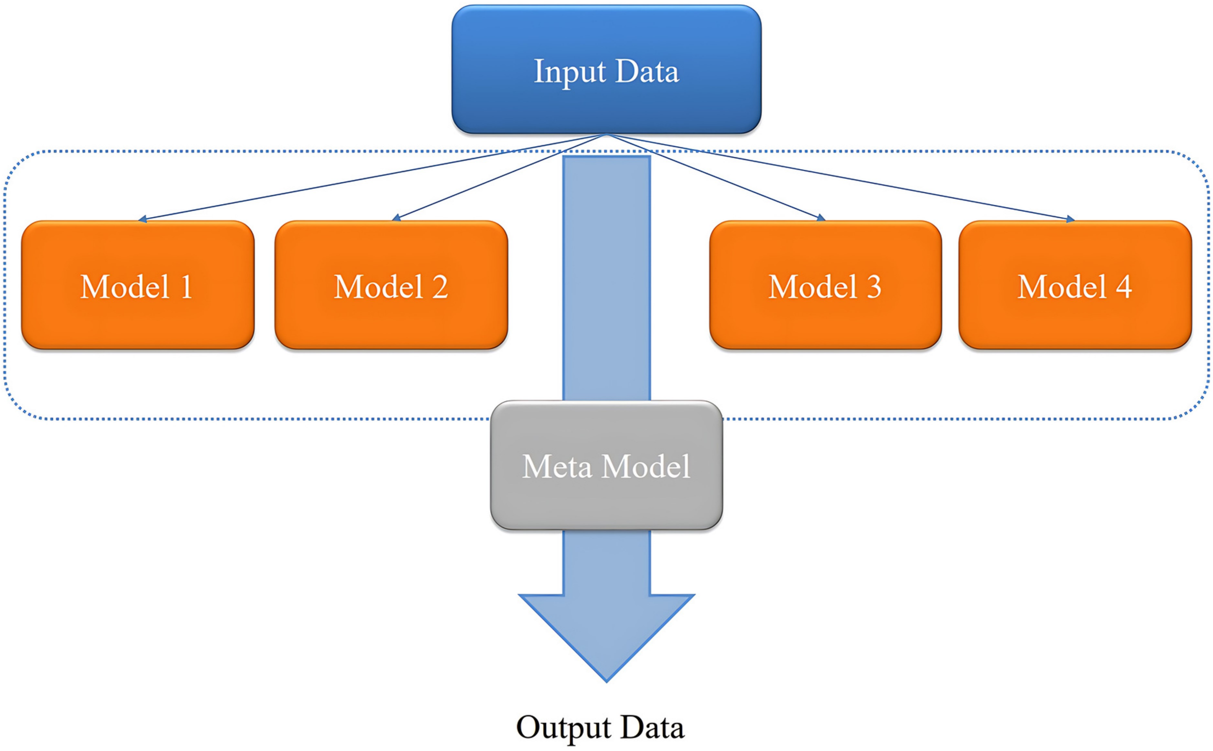

FIGURE 3. Stacking ensemble visualization. The data will be processed by each base model in the ensemble. The predictions from these models will then be fed into the meta-model, which will make the final classification.

single neuron within the FFN.

○ Residual dropout randomly drops out residual connections within the residual blocks of the FT-Transformer. This helps prevent the model from overfitting to the identity mapping and encourages it to learn more complex and generalizable features.

The specific dropout rates used for each technique are likely detailed in **Supplementary Table 1**. To further mitigate overfitting, we likely employed a separate hold-out test set during model training. This dataset was not used for training the model but served solely for evaluating its generalizability on unseen data. The performance metrics (accuracy, AUC score, *etc.*) obtained on the hold-out test set provided a more realistic assessment of the model's ability to perform well on new data, ultimately preventing overfitting to the training data.

3.5 Performance metrics

A set of performance metrics encompassing accuracy, precision, recall and F1-score was employed to evaluate the effectiveness of the prediction models was evaluated. These metrics are computed based on the following formulas:

$$Accuracy = \frac{(True_{positive} + True_{negative})}{(True_{positive} + True_{negative} + False_{positive} + False_{negative})} \tag{3}$$

$$Precision = \frac{True_{positive}}{(True_{positive} + False_{positive})} \tag{4}$$

$$Recall = \frac{True_{positive}}{(True_{positive} + False_{negative})} \tag{5}$$

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

Where $True_{negative}$ and $True_{positive}$ represent correct predictions for non-diabetic and diabetic patients, respectively. $False_{negative}$ and $False_{positive}$ indicate incorrect predictions for these groups.

Supplementary to the performance metrics previously discussed, we utilized the "Area Under the Receiver Operating Characteristic Curve" (AUC) as a pivotal yardstick for evaluating the prowess of our prediction models. The AUC score stands as a sturdy performance gauge that is not influenced by specific classification thresholds [47]. A higher AUC value signifies an elevated predictive aptitude of a model. Throughout our investigation, we designated the model attaining the peak AUC value as harboring the most exceptional predictive capacity. In instances where multiple models achieved identical AUC values, preeminence was accorded to the model boasting the highest F1-score. To execute our research endeavors, we harnessed the capabilities of Jupyter Notebook in tandem with Python 3.11.4. In particular, we leveraged the "sklearn" library to design various ML models such as CatBoost, LightGBM, XGBoost and RF, while the "pytorch" package played a pivotal role in realizing the TabNet model and FT-Transformer model.

3.6 SHapley Additive exPlanations

We employed SHAP (SHapley Additive exPlanations) [23] to gain insights into the feature importance and contribution to model predictions for men's diabetes classification. SHAP considers every possible combination of features to assess how much each feature contributes to the final prediction [48]. SHAP calculates the contribution of each feature value to the model's prediction by considering all possible combinations of features. This contribution is then weighted and summed up to arrive at a Shapley value:

$$\phi_j(val) = \textstyle\sum_{S\subset\{1,\dots,p\}\{j\}} \frac{|S|!\,(p-|S|-1)!}{p!}\,(val\,(S\cup\{j\}-val\,(S))) \tag{7}$$

This calculation considers all possible combinations of features (represented by *S*) in the model. It then focuses on a specific data point (instance to be explained) with its own set of feature values (represented by the vector *x*). There are *p* total features in the model. The part represents the model's prediction when only considering the features in set *S*, but accounting for the average effect of all other features:

$$val_x\,(S) \;=\; \int \hat{f}(x_1,\dots,x_p)d\mathbb{P}_{x\notin S} - E_X(\hat{f}\,(X)) \tag{8}$$

To understand the nuanced effects of features on individual predictions, a SHAP Beeswarm Plot was employed. Each data point was represented by a single point on the plot, positioned along the x-axis according to its SHAP value for a specific feature. The density of points in each feature row indicated the strength of that feature's influence on the model's prediction for that specific feature. Finally, we utilized SHAP Local Waterfall Plots to deconstruct the model's prediction for individual data points. This visualization commenced with the baseline prediction (average prediction on the training set) and sequentially displayed how each feature value in that data point either increased (red) or decreased (blue) the prediction. The SHAP explainer functions were implemented from the SHAP Python module by Slundberg *et al.* [49] available at https://github.com/slundberg/shap.

4. Results

4.1 Results of feature selection process

We employed the BorutaShap technique to identify the most important features for the default configurations of RF, LightGBM, CatBoost and XGBoost models. This analysis revealed 76, 44, 61 and 37 significant features for each model, respectively. Details regarding the evaluation of feature selection methods for these base models using their default hyperparameters are presented in Table 2.

When all features were included, the FT-Transformer model stood out with the highest accuracy (0.8625) and second-highest precision (0.8063) among all methods. It also performed well in other metrics, achieving a recall of 0.7371,

TABLE 2. Performance of feature selection techniques on male diabetes classification data.

Feature selection area	Model	Accuracy	Precision	Recall	F1-score	AUC
Using all feature						
	RF	0.8446	0.7389	0.8152	0.7752	0.8371
	LightGBM	0.8554	0.7725	0.7935	0.7828	0.8396
	CatBoost	0.8607	0.7849	0.7935	0.7892	0.8435
	XGBoost	0.8571	0.7708	0.8043	0.7872	0.8437
	TabNet	0.7893	0.6435	0.8043	0.715	0.7931
	FTT	0.8625	**0.8063**	0.7371	0.7701	0.8283
RF Features						
	RF	0.8536	0.7406	0.8533	0.7929	0.8535
	LightGBM	0.8571	0.7737	0.7989	0.7861	0.8423
	CatBoost	0.8518	0.7617	0.7989	0.7798	0.8383
	XGBoost	0.8571	0.7766	0.7935	0.7849	0.8409
	TabNet	0.8196	0.6861	0.8315	0.7518	0.8227
	FTT	0.8304	0.7597	0.6686	0.7112	0.7862
LightGBM Features						
	RF	**0.8661**	0.7711	**0.8424**	**0.8052**	**0.8600**
	LightGBM	0.8518	0.7617	0.7989	0.7798	0.8383
	CatBoost	0.8194	0.6845	0.7516	0.7165	0.8003
	XGBoost	0.8607	0.7732	0.8152	0.7937	0.8491
	TabNet	0.7804	0.6473	0.7283	0.6854	0.7671
	FTT	0.8393	0.7903	0.6405	0.7076	0.7832
CatBoost Features						
	RF	0.8482	0.7415	0.8261	0.7815	0.8426
	LightGBM	0.8554	0.7641	0.8098	0.7863	0.8437
	CatBoost	0.8554	0.7725	0.7935	0.7828	0.8396
	XGBoost	0.8429	0.7581	0.7663	0.7622	0.8233
	TabNet	0.8143	0.6942	0.7772	0.7333	0.8048
	FTT	0.8304	0.7597	0.6686	0.7112	0.7862
XGBoost Features						
	RF	0.8571	0.7653	0.8152	0.7895	0.8464
	LightGBM	0.8482	0.7735	0.7609	0.7671	0.8259
	CatBoost	0.8536	0.7713	0.788	0.7796	0.8368
	XGBoost	0.8482	0.7735	0.7609	0.7671	0.8259
	TabNet	0.8054	0.6531	0.8696	0.7459	0.8218
	FTT	0.8500	0.7826	0.7200	0.7500	0.8145

RF Features: Features selected by BorutaSHAP + RF; LightGBM Features: Features selected by BorutaSHAP + LightGBM; CatBoost Features: Features selected by BorutaSHAP + CatBoost; XGBoost Features: Features selected by BorutaSHAP + XGBoost; XGBoost: XGBoost Classifiers; CatBoost: CatBoost Classifier; LightGBM: Light Gradient Boosting Classifier; TabNet: TabNet Classifier; FTT: (Feature Tokenizer + Transformer) Classifier; RF: Random Forest Classifier; AUC: Area Under the Receiver Operating Characteristic Curve. The bolded numbers represent the highest value for each metric.

F1-score of 0.7701, and AUC of 0.8283. CatBoost followed closely with an accuracy of 0.8607 and AUC of 0.8435. The remaining models ranked as XGBoost, LightGBM, RF and TabNet.

However, the performance landscape shifted significantly after incorporating features identified as important by BorutaShap with LightGBM. Surprisingly, RF emerged as the top performer across all metrics except precision. It achieved scores of 0.8661 for accuracy, 0.8424 for recall, 0.8052 for F1-score, and a remarkable 0.8600 for AUC. Both FTT and

TabNet models saw a drop in performance when using the selected features. Their AUC values decreased to 0.7832 and 0.7671, respectively.

The three models that utilized BorutaShap feature selection (XGBoost, CatBoost and RF) consistently outperformed the two DL models (FTT and TabNet). Furthermore, TabNet's performance even dipped below FTT when using the selected features. This suggests that, in this specific case, the DL models might benefit more from leveraging the full set of available features, while traditional ML models can achieve better results with a carefully chosen subset.

Based on these findings, we selected the subset of features chosen by LightGBM's BorutaShap technique for optimal performance. This choice was driven by RF's superior performance across all evaluation metrics. Details of these features are provided in Table 3, where □□□ represents the type of input number.

4.2 Performance comparison among optimized models

After fine-tuning with Optuna, the optimal hyperparameters for each model are presented in **Supplementary Table 1**. Analyzing the effectiveness of fine-tuning on our base models for predicting individuals with diabetes reveals valuable insights (Table 4). LightGBM achieved the highest AUC score (0.8706) with a recall of 0.8686 and F1-score of 0.8085. Notably, the FT-Transformer model exhibited a different strength, surpassing LightGBM in accuracy (0.8732) and precision (0.8095). Comparing these results with those in Table 2, we observe improvements in AUC for LightGBM (0.8383 to 0.8706) and TabNet (0.7804 to 0.8482). However, the impact is not uniform. Random Forest's performance decreased in both accuracy and AUC after fine-tuning, while XGBoost and CatBoost showed gains.

Following an analysis of model performance, we designed a stacking ensemble model that capitalizes on the strengths of four specific models. While F1-score is a crucial metric for balanced assessment, we considered a broader range of performance indicators during model selection for the ensemble. FT-Transformer's exceptional accuracy and precision make it ideal for capturing highly accurate predictions within the ensemble. LightGBM, on the other hand, brings robust and balanced predictions to the table thanks to its leading performance in recall, F1-score and AUC. XGBoost strengthens the overall performance with its second-highest overall AUC. Finally, RF, while not the top performer in all metrics, contributes valuable diversity (third-highest AUC) which can improve the generalizability of the final ensemble model. By combining these diverse models, we aim to create a stacking ensemble that surpasses the individual performance of any single model.

Table 5 summarizes the performance of three stacking ensemble models. Notably, the model using Random Forest (RF) as the meta-learner achieved the highest accuracy (0.8786), recall (0.8171), F1-score (0.8079) and AUC (0.8618). In contrast, the LightGBM-based model achieved the highest precision but lower performance in other metrics. Compared to the individual optimized models, our proposed stacking ensemble method surpasses them in accuracy (0.8786) and AUC (0.8616). However, it trades this improvement for slightly lower precision, recall and F1-score. This suggests the model strikes a balance between capturing accurate predictions and identifying positive cases effectively. Fig. 4 presents a comparison of the Receiver Operating Characteristic (ROC) curves for the stacking model and the standalone models. The confusion matrix for the RF meta-classifier ensemble with FT-Transformer, XGBoost and LightGBM is shown in Fig. 5.

4.3 Interpretation the proposed stacking ensemble model with SHAP

Our top-performing stacking ensemble model, which utilizes RF as the meta-learner, highlights the significance of various features in predicting diabetes, as illustrated in Fig. 6. Notably, waist circumference (HE_WC) emerges as a crucial predictor. Existing research indicates a strong correlation between larger waist circumference and increased diabetes risk, particularly among individuals with a lower BMI [50]. Age is another prominent feature, with a general trend showing that the risk of developing Type 2 diabetes escalates with advancing age. Additionally, urine biomarkers such as elevated levels of blood creatinine (HE_UCREA), uric acid (HE_UACID), and urine albumin (HE_UALB) are associated with kidney disease, a known risk factor for Type 2 diabetes [51, 52].

To gain a deeper understanding of the model's decision-making process, we examined three specific individuals with varying predicted outcomes (Figs. 7,8,9). We leveraged SHAP's color-coded visualizations to pinpoint features that significantly influenced the model's prediction for each person. Here, red hues highlight features that strongly support a Class 1 prediction (indicating diabetes). Conversely, features aligned with a Class 0 prediction (non-diabetic) are displayed in blue. The first individual's attributes are as follows:

- HE_UALB: 99.199997 μg/mL
- HE_WC: Waist circumference 122.199997 cm
- HE_NC: Neck circumference 44.5 cm
- HE_UACID: Uric acid 4.2 mg/dL
- HE_DMFH1: Father got diabetes (1/Yes)
- DI1_DG: Doctor diagnosed high blood pressure? (1/Yes)
- HE_WBC: White blood cell count 7.81 thousands/μL
- Age: 43
- HE_ALT: Alanine Aminotransferase (ALT) Test (SGPT) 43.0 International Unit (IU)/L
- AGE: 34
- HE_UCREA: Blood creatinine 211.300003 mg/dL
- HE_NC: Neck circumference N/A
- HE_UACID: Uric acid 6.7 mg/dL
- HE_WBC: White blood cell count 5.28 thousands/μL
- DI1_DG: Doctor diagnosed high blood pressure? (0/No)
- HE_CHOL: Total cholesterol 217.0 mg/dL
- N_DIET_WHY: Reasons for diet therapy? (2/To control weight)
- HE_WC: Waist circumference 89.0 cm
- HE_WC: Waist circumference 79.800003 cm
- HE_UACID: Uric acid 2.9 mg/dL
- HE_TG: Neutral fat 72.0 mg/dL
- N_DIET_WHY: Reasons for diet therapy: 8/non-glycemic (dietary therapy: no)

TABLE 3. Description of selected feature used in this research.

Variable	Description	Type	Values and description
DI2_DG	Dyslipidemia Physician Diagnosed	Categorical	Dyslipidemia diagnosed by doctor? 0. None 1. Yes 8. Not eligible (children, adolescents) 9. Don’t know, no response
BP5	Depression for more than two weeks in a row	Categorical	Depressed for more than 2 weeks in a row 1. Yes 2. No 8. Not applicable (adults, children) 9. Don’t know, no response
BM1_8	When to Brush Your Brushes: Before You Go to Bed	Categorical	When to brush teeth: Before going to bed 0. No 1. Yes 8. Not applicable (did not brush teeth) 9. I don’t know
HE_UACID	Uric acid	Continuous	Uric acid □□.□ mg/dL
HE_WC	Waist circumference	Continuous	Waist circumference □□□.□ cm
EC_STT_2	Occupational status: Details of wage workers	Categorical	Occupational status: wage earner details 1. Full-time employee 2. Temporary work 3. Daily laborer 8. Not applicable (Question 3—②③④⑧) 9. Don’t know, no response
BM8	A speaking question	Categorical	Speech problems 1. Very uncomfortable 2. Inconvenience 3. So-so 4. Not uncomfortable 5. Not uncomfortable at all 8. Not applicable (under 19 years old) 9. I don’t know
HE_WBC	White blood cell count	Continuous	White blood cell count □□.□□ Thousands/μL
N_EN	Energy intake (Kcal)	Continuous	Daily energy intake (kcal)
HE_TG	Neutral fat	Continuous	Neutral fat □□□□ mg/dL
HE_ALT	ALT (SGPT)	Continuous	Alanine Aminotransferase (ALT) Test (SGPT) □□□ IU/L
BP17_DG	A doctor’s diagnosis of obstructive sleep apnea	Categorical	Doctor diagnosed with obstructive sleep apnea? 0. No 1. Yes 8. Not applicable (under 40 years old) 9. Don’t know, no response
N_DIET	Meal therapy status	Categorical	Whether it is diet therapy or not? 1. Yes 2. No 9. Don’t know/No response
HO_INCM	Income Quaternary (household)	Categorical	Income quartile (household) Refer to the standard amount for 4th quartile classification. 1. High 2. Low-mid 3. Low 4. Award
BM7	A chewing question	Categorical	Chewing Problems 1. Very uncomfortable 2. Inconvenience 3. So-so 4. Not uncomfortable 5. Not uncomfortable at all 8. Not applicable (under 19 years old) 9. “I don’t know”

TABLE 3. Continued.

Variable	Description	Type	Values and description
TINS	Type of health insurance	Continuous	Types of health insurance 10. National Health Insurance (regional) 20. National Health Insurance (workplace) 30. Medical benefits 99. Not registered, don’t know, no response
HE_HT	Kidney	Continuous	Height □□□.□ cm
HE_DMFH3	Diabetes doctor diagnosis (siblings)	Categorical	Diabetes diagnosed by a doctor (siblings) 0. No 1. Yes 8. Not applicable 9. Don’t know/No response
HE_CHOL	Total cholesterol	Continuous	Total cholesterol □□□ mg/dL
DI2_2	dosing of dyslipidemia	Categorical	Take medication for dyslipidemia 1. Take it daily 2. Taken more than 20 days a month 3. Taken more than 15 days a month 4. Taken less than 15 days a month 5. Do not take it 8. Not applicable(Children, adolescents, not diagnosed by a doctor) 9. Don’t know, no response
N_DIET_WHY	Reasons for dietary therapy	Categorical	Reasons for diet therapy 1. Having a disease 2. To control weight 3. Others 8. Non-glycemic (dietary therapy: no) 9. Don’t know/No response
AGE	American age	Continuous	□□ years old
DI1_DG	Whether to be diagnosed as a doctor with high blood pressure	Categorical	Doctor diagnosed high blood pressure? 0. None 1. Yes 8. Not eligible (children, adolescents) 9. Don’t know, no response
HE_NC	Neck circumference	Continuous	Neck circumference □□□.□ cm
BD7_4	Family/doctor’s recommendation to abstain from drinking	Categorical	Whether or not your family/doctor recommends abstinence? 1. There was no 2. It existed in the past, but not in the past year. 3. It happened in the last year 8. Not applicable (Question 1—①, children, adolescents) 9. Don’t know, no response
DI1_PR	Current prevalence of high blood pressure	Categorical	Current presence of high blood pressure? 0. None 1. Yes 8. Not applicable(Children, adolescents, not diagnosed by a doctor) 9. Don’t know, no response
MARRI_2	Marital status	Categorical	Married? 1. Spouse, living together 2. Existing spouse, separation 3. Bereavement 4. Divorce 8. Refusal to respond 9. Don’t know 88. Not applicable (Question 10—②) 99. No response
DI2_PR	Dyslipidemia Current prevalence	Categorical	Current presence of dyslipidemia? 0. None 1. Yes 8. Not applicable(Children, adolescents, not diagnosed by a doctor) 9. Don’t know, no response

TABLE 3. Continued.

Variable	Description	Type	Values and description
BO1_2	The amount of weight loss in a year	Categorical	Weight loss in 1 year 1. More than 3 kg–less than 6 kg 2. More than 6 kg–less than 10 kg 3. Over 10 kg 8. Not applicable (Question 2—①③⑧) 9. Don't know, no response
BO1_1	Weight change in 1 year	Categorical	Subjective body shape recognition 1. Very skinny 2. A bit thin 3. Normal 4. Slightly obese 5. Very obese 8. Not applicable (under 6 years old) 9. Don't know, no response
HE_BUN	Blood urea nitrogen	Continuous	Blood urea nitrogen □□ mg/dL
HE_DMFH1	Diabetes doctor diagnosis status (father)	Categorical	Diabetes diagnosed by a doctor (father) 0. No 1. Yes 9. Don't know/No response
HE_UALB	Urinary albumin	Continuous	Urinary albumin □□□□.□ μg/mL
N_WAT_C	Water intake (cup)	Continuous	Water intake (cup: 200 mL)
BM1_4	Time to brush: After lunch	Continuous	When to brush teeth: After lunch 0. No 1. Yes 8. Not applicable (did not brush teeth) 9. I don't know
HE_UCREA	Heavy creatinine	Continuous	Blood creatinine □.□□ mg/dL
BO1_3	The amount of weight gain in a year	Categorical	Weight gain in 1 year
HE_HB	Hemoglobin	Continuous	Hemoglobin □□.□ g/dL
LF_SAFE	The dietary situation of the past year	Categorical	Eating situation 1. Being able to eat a sufficient amount and variety of food there was. 2. I was able to eat a sufficient amount of food, but I couldn't eat a variety of foods. 3. Sometimes food is difficult due to financial difficulties. It wasn't enough. 4. It is difficult to eat often due to financial difficulties. It wasn't enough. 9. Don't know/No response
L_BR	Whether to skip breakfast one day before the food intake survey	Categorical	2 days before food intake survey Skipping breakfast or not 1. Yes 0. No
HE_FST	An empty stomach	Continuous	Fasting time □□ time
HE_BPLT	Platelet count	Continuous	Platelet count □□□ Thousands/μL
HE_AST	AST (SGOT)	Continuous	Aspartate transaminase (AST) (SGOT) □□□ IU/L
HE_UPH	Uric acidity	Continuous	Uric Acid Level □.□

□□□ represents the type of input number; SGOT: serum glutamic oxaloacetic transaminase; AST: aspartate aminotransferase; SGPT: serum glutamate pyruvate transaminase; ALT: alanine transaminase; IU: International Unit.

TABLE 4. Diabetic prediction performance of optimized models.

Model	Accuracy	Precision	Recall	F1-score	AUC
RF	0.8446	0.7075	0.8571	0.7752	0.8496
LightGBM	0.8714	0.7562	**0.8686**	**0.8085**	**0.8602**
CatBoost	0.8554	0.7374	0.8343	0.7828	0.8481
XGBoost	0.8625	0.7552	0.8286	0.7902	0.8532
TabNet	0.8482	0.7813	0.7143	0.7463	0.8217
FTT	**0.8732**	**0.8095**	0.7771	0.7930	0.8470

XGBoost: XGBoost Classifier; CatBoost: CatBoost Classifier; LightGBM: Light Gradient Boosting Classifier; TabNet; TabNet Classifier; FTT: (Feature Tokenizer + Transformer) Classifier; RF: Random Forest Classifier; AUC: Area Under the Receiver Operating Characteristic Curve. The bolded numbers represent the highest value for each metric.

TABLE 5. Three stacking models' performance.

Meta model	Accuracy	Precision	Recall	F1-score	AUC
RF	**0.8786**	0.7989	**0.8171**	**0.8079**	**0.8618**
LightGBM	0.8625	**0.8141**	0.7257	0.7674	0.8252
XGB	0.8696	0.7931	0.7886	0.7908	0.8475

RF: RF Classifier as meta model + (FT-Transformer + XGBoost + LightGBM); LightGBM: LightGBM Classifier as meta model + (FT-Transformer + XGBoost + RF); XGB: XGBoost Classifier as meta model + (FT-Transformer + LightGBM + RF); AUC: Area Under the Receiver Operating Characteristic Curve. The bolded numbers represent the highest value for each metric.

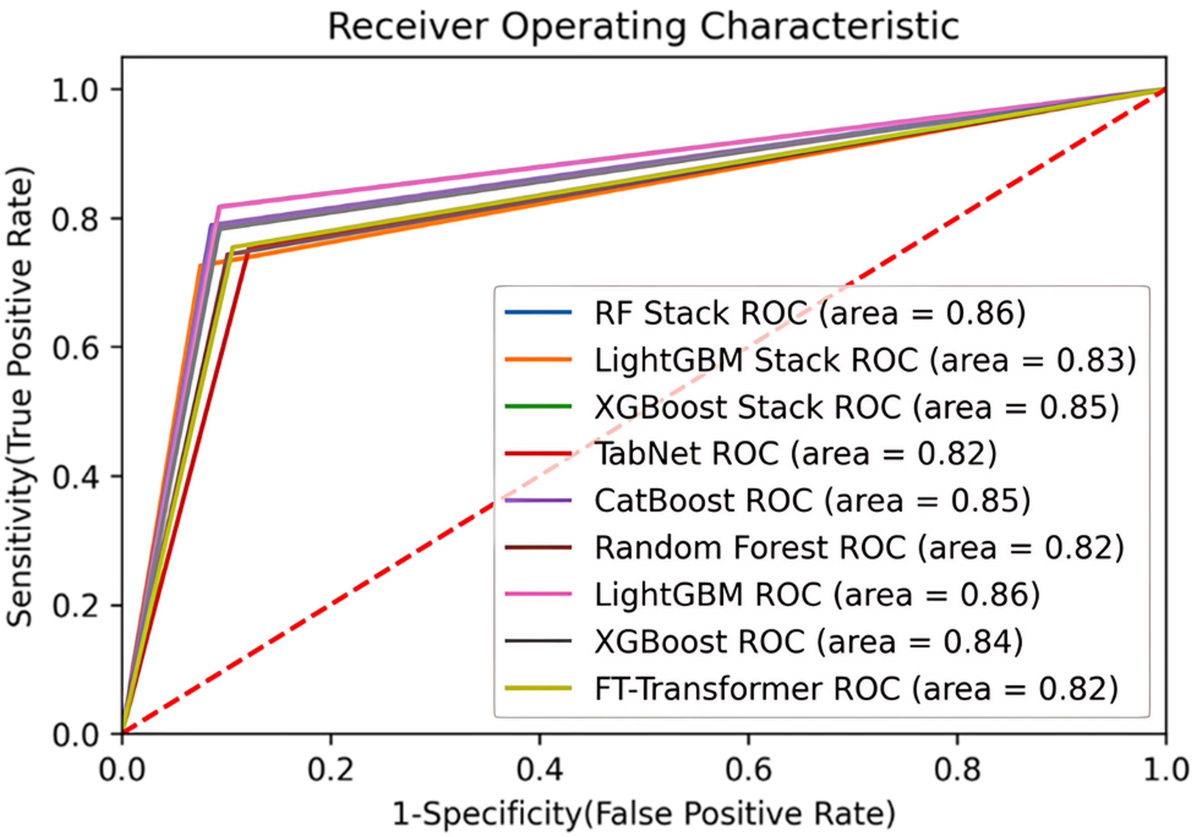

FIGURE 4. ROC of all models on the hold-out set. True Positive Rate is on the y-axis against the False Positive Rate on the x-axis. RF: Random Forest Classifier; ROC: Receiver Operating Characteristic; FT: Feature Tokenizer Transformer Classifier.

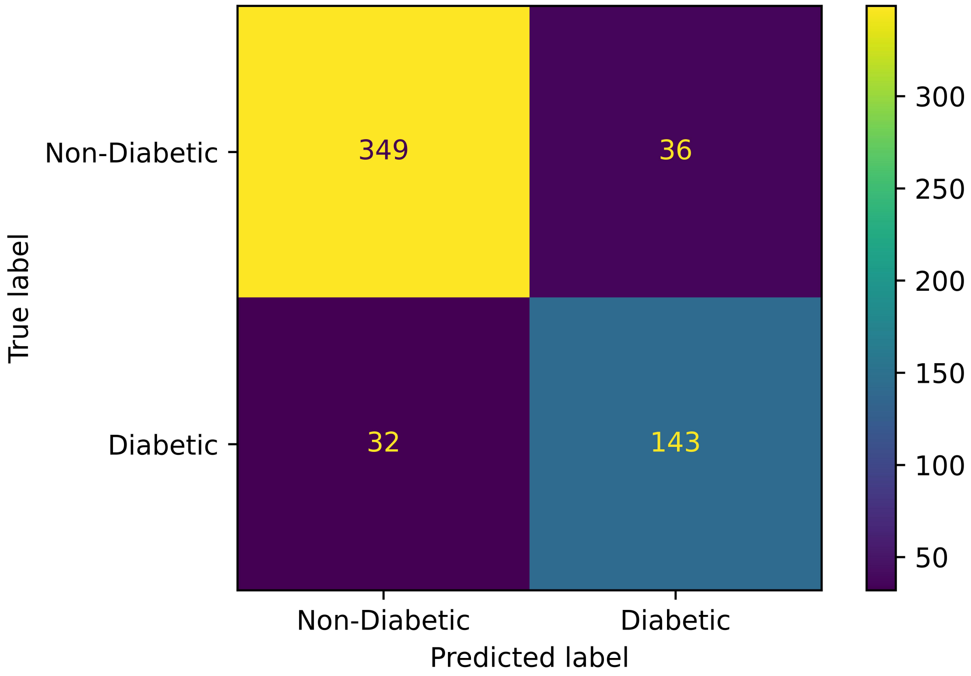

FIGURE 5. Confusion matrix of the proposed stacking model on the hold-out set. Rows represent the actual labels (True Diabetic, True Non-Diabetic, False Diabetic, False Non-Diabetic). Columns represent the predicted labels (Predicted Diabetic, Predicted Non-Diabetic). The diagonal elements (True Diabetic and True Non-Diabetic) represent correct predictions. Off-diagonal elements (False Diabetic and False Non-Diabetic) represent misclassifications.

- HE_NC: Neck circumference 35.5 cm
- HE_CHOL: Total cholesterol 181.0 mg/dL
- DI1_DG: Doctor diagnosed high blood pressure? (0/No)
- HE_AST: Aspartate transaminase (AST) (SGOT) 13.0 IU/L
- N_WAT_C: Water intake 3 cup (cup: 200 mL)

5. Discussion

This study aimed to identify an effective model for predicting men's diabetes in South Korea, specifically focusing on improving men's health outcomes. We meticulously evaluated six individual models (FT-Transformer, TabNet, LightGBM, Random Forest, XGBoost and CatBoost) along with three ensemble models created by stacking the top four performing models with different meta-models. Our findings concurred with previous research, demonstrating that the stacking ensemble models achieved superior performance compared to single-based models. Furthermore, we integrated SHAP with this ensemble model to gain valuable insights into feature importance. This approach offers a beneficial tool for both healthcare professionals experiencing burnout with traditional Electronic Health Record (EHR) analysis [7] and data analyst specialists. SHAP provides a two-fold benefit: its global explanations enable analysis of the overall importance of each feature, while its individual explanations facilitate understanding the rationale behind specific model predictions. Finally, our research highlights the critical role of feature selection methods in optimizing both accuracy and computational efficiency, ultimately leading to better model performance. A deeper exploration of feature selection techniques specifically tailored to the KNHANES dataset holds promise for further improvements. Additionally, inspired by the work of Dharmarathne *et al.* [35], the development of a user-friendly interface system could be a valuable avenue for future research. This aligns with the growing emphasis on improving model interpretability and user adoption in clinical practice.

This study investigated the application of an ensemble model for men's diabetes prediction using the 7th KNHANES dataset (2016–2018). This approach yielded promising results, achieving an accuracy of 0.8786, recall of 0.8171, F1-score of 0.8079 and AUC of 0.8618. Here, we compare these findings with relevant studies to contextualize the performance of our proposed model.

Chang *et al.* [30] utilized an ANN model on the same 7th KNHANES dataset, achieving an accuracy of 0.813. While both studies achieved high accuracy on this dataset, our ensemble model demonstrates a slight improvement, suggesting its potential effectiveness for diabetes prediction in this specific population. In contrast, Choi *et al.* [31] employed a SVM model on the 5th KNHANES data, attaining an AUC of 0.731 and an accuracy of 0.661 for prediabetes prediction. It's important to note that our study focused on confirmed diabetes cases, and the achieved higher AUC suggests better performance in distinguishing diabetic from non-diabetic patients. Joseph *et al.* [8] reported an F1-score of 0.88 using a Bayesian TabNet model on the Pima Indians Diabetes Dataset, which is a smaller and more traditional benchmark dataset compared to the 7th

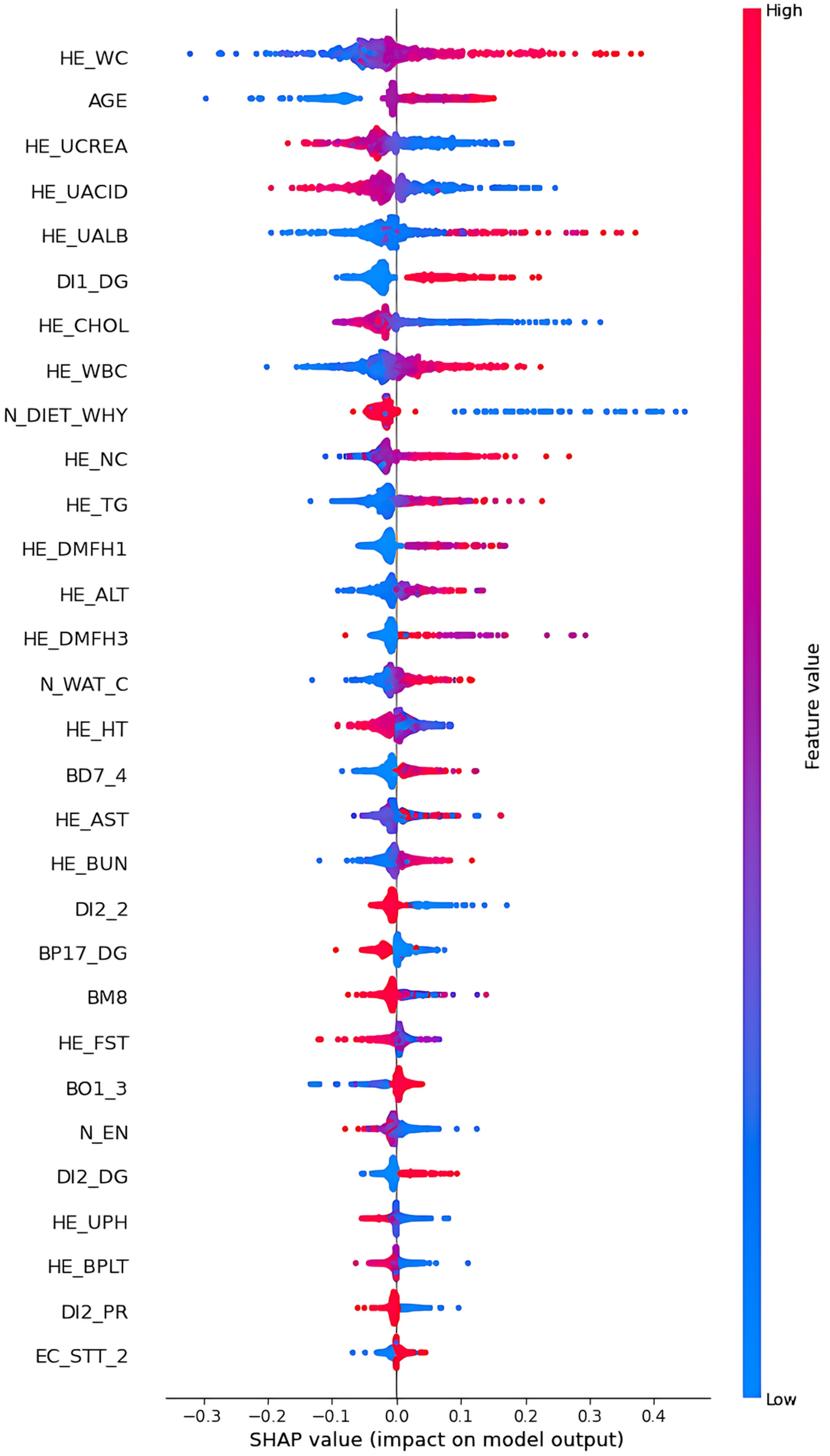

FIGURE 6. SHAP's global explanation. Beeswarm plot showing how relevant each feature is. SHAP: SHapley Additive exPlanations; The feature name used in Fig. 6 can be found in Table 3.

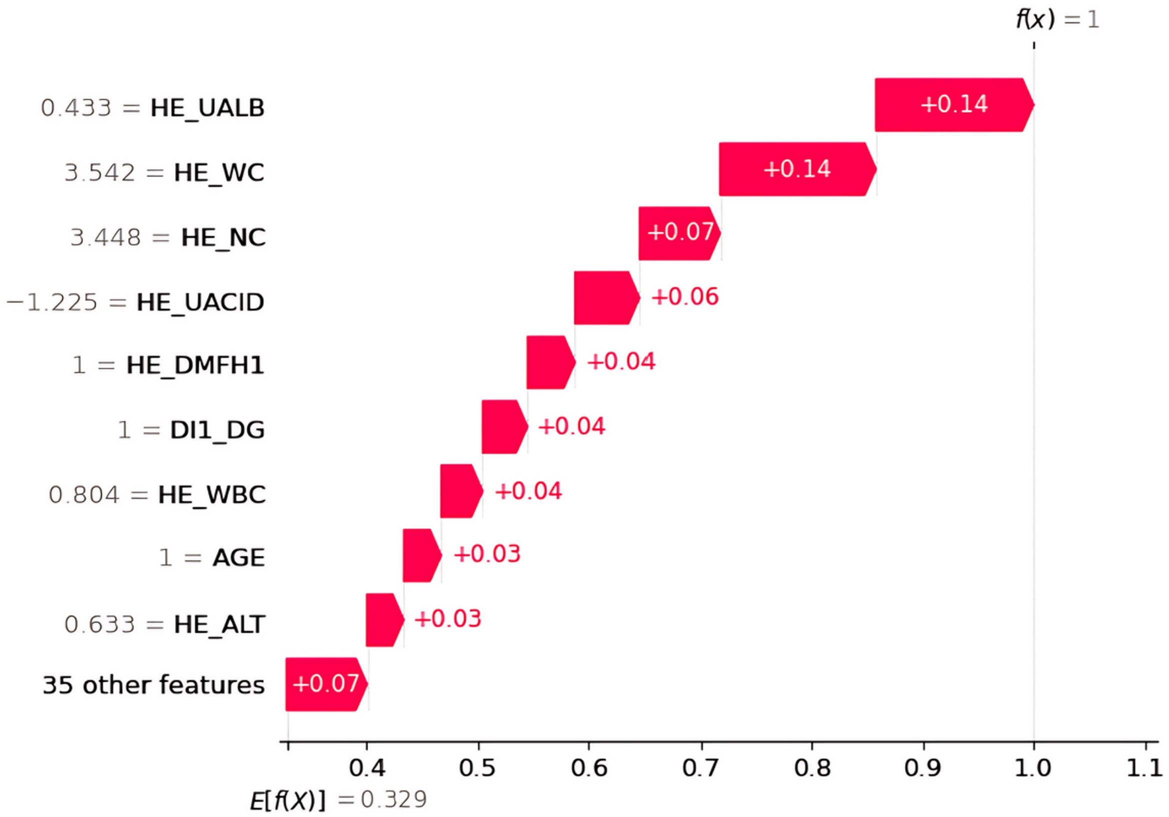

FIGURE 7. SHAP's local explanation. The initial instance's prediction explained by waterfall plot. The model predicted Class 0 (non-diabetic) for the fourth case, indicated by the blue color alignment in the SHAP visualization.

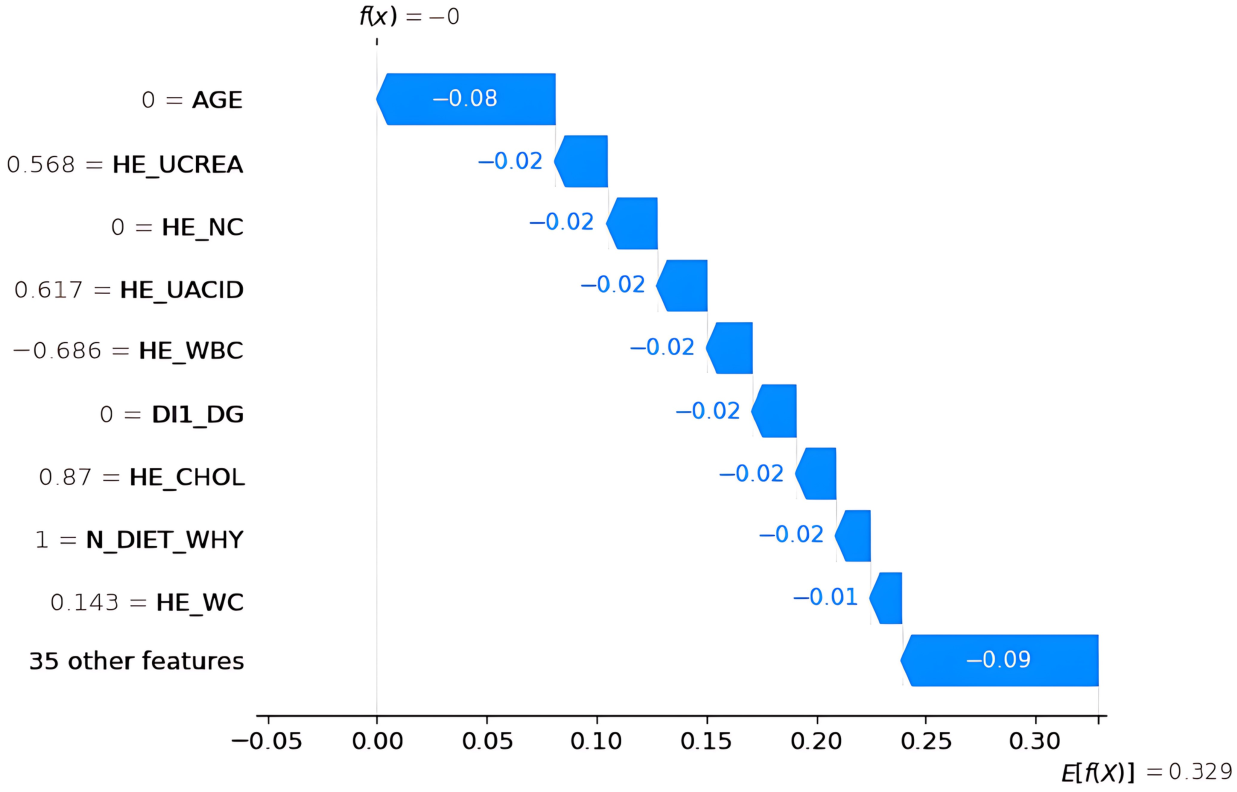

FIGURE 8. SHAP's local explanation. The fourth instance's prediction explained by waterfall plot. The SHAP visualization for the fifth prediction showed a mix of red and blue features, indicating some influence from both Class 1 (diabetic) and Class 0 (non-diabetic) features. However, the model ultimately classified this case as Class 0.

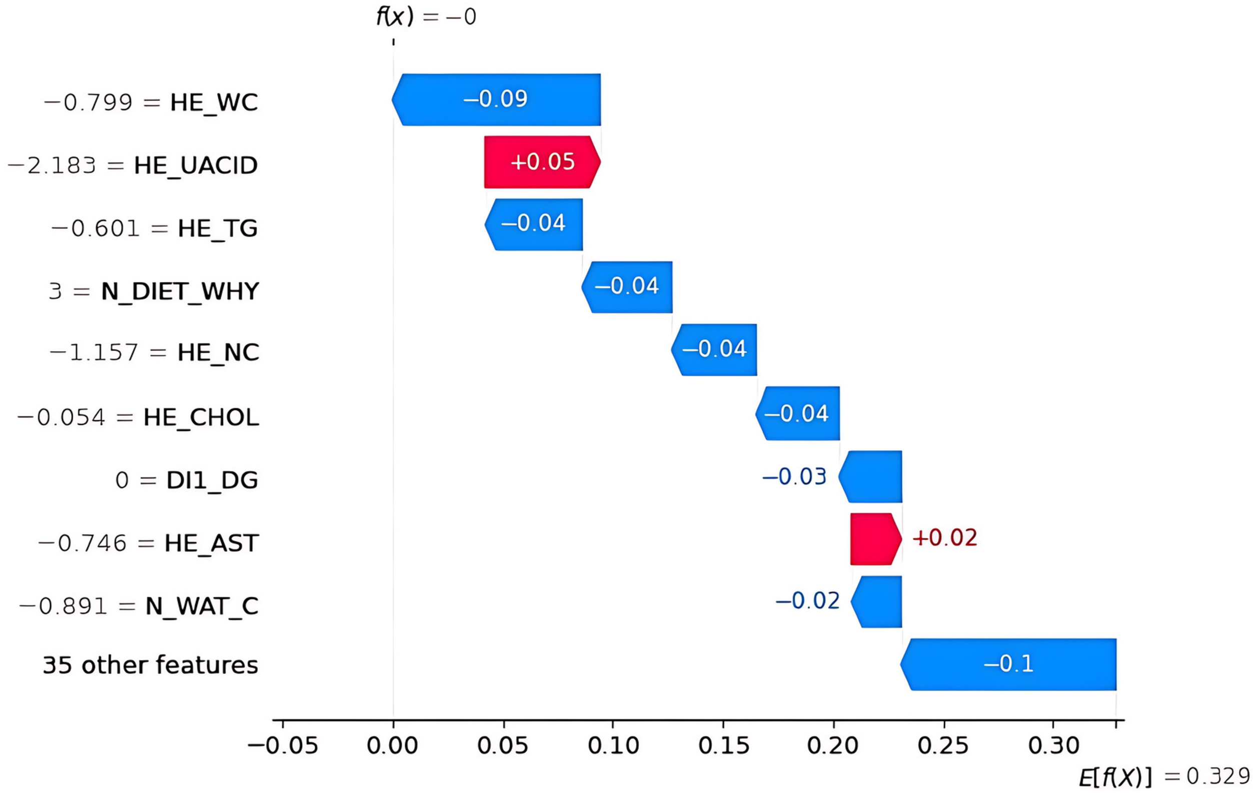

FIGURE 9. SHAP's local explanation. The fifth instance's prediction explained by waterfall plot.

KNHANES data used in our study. The inherent differences in dataset complexity can influence model performance, potentially explaining the slight variation in TabNet performance between the two studies.

Our study highlights the potential of DL techniques, particularly deep neural networks like the FT-Transformer specifically designed for tabular data, and ensemble DL methods for future research in diabetes prediction. The strong performance of the FT-Transformer model using all features merits further investigation to understand the underlying factors contributing to its success. Given the limitations in existing research on men's diabetes health, further investigation is warranted to bridge this knowledge gap. This exploration holds the potential to not only advance the development of generalizable DL models for diabetes prediction, but also lead to the creation of more specific and effective models tailored for the male population.

Our study acknowledges some limitations. While focusing on men fills a gap in existing research, it limits the generalizability of our findings. Expanding future studies to include women and other demographic groups would provide a more holistic understanding of the model's performance across diverse populations.

Furthermore, SHAP offers valuable insights into the internal workings of the model by explaining individual feature attributions, but it may not fully satisfy the interpretability needs of clinicians in a real-world setting. In clinical decision-making, achieving a level of causal interpretability is crucial [53]. Causability, in the context of clinical practice, goes beyond simply knowing which features are important and delves into understanding why they are important and how they influence the model's output. This deeper understanding considers factors like the effectiveness of the model's predictions for patient outcomes, the efficiency of its use in clinical workflow, and user satisfaction amongst healthcare professionals. To address this limitation and enhance the model's interpretability for clinical use, future research will explore methods that combine SHAP with expert explanations. This combined approach holds promise for providing clinicians with a more comprehensive understanding of the model's reasoning process and promoting informed clinical decision-making [54].

Our application of SMOTE addressed the class imbalance in the men' diabetes dataset, potentially improving model evaluation and performance. However, limitations exist, including overfitting due to potentially unrealistic synthetic data, dependence on data quality, and potential bias towards the oversampled class. Future work exploring alternative oversampling techniques (SMOTE and Edited Nearest Neighbor's (SMOTEENN), Adaptive synthetic sampling (ADASYN), Borderline-SMOTE) could mitigate these limitations. Moreover, while the proposed model achieves high accuracy and AUC score, its F1-score, precision, and recall fall below those of standalone models. Further investigation is necessary to understand and potentially improve these metrics.

6. Conclusions

This study aimed to develop a novel method for predicting diabetes in men using a tabular dataset. We compared the performance of DL, ML and ensemble methods. Consistent with previous research, DL models exhibited slightly lower performance compared to ML models. This suggests that DL might be better suited for tasks with a vast number of features, while ML excels with well-chosen features. To leverage the strengths of both approaches, we proposed a stacking ensemble method that incorporates a Feature Tokenizer and a conventional Gradient Boosted Decision Trees (GBDT) method. Due to the model's complexity, we employed SHAP to enhance the interpretability of the predictions. Future research could delve deeper into this domain using larger datasets. Additionally, developing a user-friendly interface and addressing challenges like missing data, class imbalance and feature importance could improve the model's trustworthiness and pave the way for its application in clinical settings, particularly for men's health.

AVAILABILITY OF DATA AND MATERIALS

This study leveraged data from the Korea National Health and Nutrition Examination Survey (KNHANES) conducted between 2019 and 2022. Data are publicly available through the KNHANES website. More information at: http://knhanes.cdc.go.kr.

AUTHOR CONTRIBUTIONS

VQT and HB—designed the research study; wrote the manuscript. VQT—performed the research; analyzed the data. HB—provided help and advice on visualization. YC and HB—were the supervisor of this study and provided a fund for the study. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Before the survey, written informed consent was obtained from each study participant. The current study used only existing de-identified data. The study was conducted according to the guidelines of the Declaration of Helsinki. The protocol of 2016–2018 KNHANES was approved by the Institutional Review Board (IRB) of the Korea Centers for Disease Control and Prevention (IRB approval number in 2016–2018: 2018-01-03-P-A and 2018-01-03-C-A). Written informed consent was obtained from all participants.

ACKNOWLEDGMENT

Thanks for the support of the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF-RS-2023-00237287, NRF-2021S1A5A8062526), and the local government-university cooperation-based regional innovation projects (2021RIS-003).

FUNDING

This research Supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-RS-2023-00237287, NRF-2021S1A5A8062526) and local government-university cooperation-based regional innovation projects (2021RIS-003).

CONFLICT OF INTEREST

The authors declare no conflict of interest. Haewon Byeon is serving as one of the Guest Editors of this journal. We declare that Haewon Byeon had no involvement in the peer review of this article and has no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to YC.

SUPPLEMENTARY MATERIAL

Supplementary material associated with this article can be found, in the online version, at https://oss.jomh.org/files/article/1862388728403509248/attachment/Supplementary%20material.docx.

REFERENCES

[1] Petersmann A, Müller-Wieland D, Müller UA, Landgraf R, Nauck M, Freckmann G, *et al*. Definition, classification and diagnosis of diabetes mellitus. Experimental and Clinical Endocrinology & Diabetes. 2019; 127: S1–S7.

[2] World Health Organization. The top 10 causes of death. 2024. Available at: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (Accessed: 25 May 2024).

[3] World Health Organization. Diabetes. 2023. Available at: https://www.who.int/news-room/fact-sheets/detail/diabetes (Accessed: 25 May 2024).

[4] Atkinson MA, Eisenbarth GS, Michels AW. Type 1 diabetes. The Lancet. 2014; 383: 69–82.

[5] Kautzky-Willer A, Leutner M, Harreiter J. Sex differences in type 2 diabetes. Diabetologia. 2023; 66: 986–1002.

[6] Kirwan JP, Sacks J, Nieuwoudt S. The essential role of exercise in the management of type 2 diabetes. Cleveland Clinic Journal of Medicine. 2017; 84: S15–S21.

[7] Budd J. Burnout related to electronic health record use in primary care. Journal of Primary Care & Community Health. 2023; 14: 21501319231166921.

[8] Joseph LP, Joseph EA, Prasad R. Explainable diabetes classification using hybrid Bayesian-optimized TabNet architecture. Computers in Biology and Medicine. 2022; 151: 106178.

[9] Wang Q, Cao W, Guo J, Ren J, Cheng Y, Davis DN. DMP_MI: an effective diabetes mellitus classification algorithm on imbalanced data with missing values. IEEE Access. 2019; 7: 102232–102238.

[10] Abdulhadi N, Al-Mousa A. Diabetes detection using machine learning classification methods. 2021 International Conference on Information Technology (ICIT). IEEE: Amman, Jordan. 2021.

[11] Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? To be published in arXiv. 2022. [Preprint].

[12] Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Information Fusion. 2022; 81: 84–90.

[13] El-Rashidy N, ElSayed NE, El-Ghamry A, Talaat FM. Utilizing fog computing and explainable deep learning techniques for gestational diabetes prediction. Neural Computing and Applications. 2023; 35: 7423–7442.

[14] Du Y, Rafferty AR, McAuliffe FM, Wei L, Mooney C. An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. Scientific Reports. 2022; 12: 1170.

[15] Khanna VV, Chadaga K, Sampathila N, Prabhu S, P RC, Bhat D, *et al*. Explainable artificial intelligence-driven gestational diabetes mellitus prediction using clinical and laboratory markers. Cogent Engineering. 2024; 11: 2330266.

[16] Gorishniy Y, Rubachev I, Khrulkov V, Babenko A. Revisiting deep learning models for tabular data. 2023. Available at: http://arxiv.org/abs/2106.11959 (Accessed: 25 May 2024).

[17] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, *et al*. Attention is all you need. 2023. Available at: http://arxiv.org/abs/1706.03762 (Accessed: 25 May 2024).

[18] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery: New York, NY, USA. 2016.

[19] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al*. LightGBM: a highly efficient gradient boosting decision tree. 2017. Available at: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html (Accessed: 25 May 2024).

[20] Breiman L. Random forests. Machine Learning. 2001; 45: 5–32.

[21] Arik SÖ, Pfister T. TabNet: attentive interpretable tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence. 2021; 35: 6679–6687.

[22] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. 2018. Available at: https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html (Accessed: 25 May 2024).

[23] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 2017. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf (Accessed: 25 May 2024).

[24] Guha A. Building explainable and interpretable model for diabetes risk prediction. International Journal of Engineering Research & Technology. 2020; 9: 1037–1042.

[25] Vakil V, Pachchigar S, Chavda C, Soni S. Explainable predictions of different machine learning algorithms used to predict early stage diabetes. 2021. Available at: https://arxiv.org/pdf/2111.09939 (Accessed: 25 May 2024).

[26] Vishwarupe V, Joshi PM, Mathias N, Maheshwari S, Mhaisalkar S, Pawar V. Explainable AI and interpretable machine learning: a case study in perspective. Procedia Computer Science. 2022; 204: 869–876.

[27] Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters. 2023; 10: 1–10.

[28] Kibria HB, Nahiduzzaman M, Goni MOF, Ahsan M, Haider J. An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. Sensors. 2022; 22: 7268.

[29] Dutta A, Hasan MK, Ahmad M, Awal MA, Islam MA, Masud M, *et al*. Early prediction of diabetes using an ensemble of machine learning models. International Journal of Environmental Research and Public Health. 2022; 19: 12378.

[30] Chang K, Yoo S, Lee S. Classification and prediction of the effects of nutritional intake on diabetes mellitus using artificial neural network sensitivity analysis: 7th Korea National Health and Nutrition Examination Survey. Nutrition Research and Practice. 2023; 17: 1255–1266.

[31] Choi SB, Kim WJ, Yoo TK, Park JS, Chung JW, Lee Y, *et al*. Screening for prediabetes using machine learning models. Computational and Mathematical Methods in Medicine. 2014; 2014: 618976.

[32] Kumari S, Kumar D, Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. International Journal of Cognitive Computing in Engineering. 2021; 2: 40–46.

[33] Pima Indians diabetes database. 1988. Available at: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (Accessed: 27 May 2024).

[34] Curia F. Explainable and transparency machine learning approach to predict diabetes develop. Health and Technology. 2023; 13: 769–780.

[35] Dharmarathne G, Jayasinghe TN, Bogahawaththa M, Meddage DPP, Rathnayake U. A novel machine learning approach for diagnosing diabetes with a self-explainable interface. Healthcare Analysis. 2024; 5: 100301.

[36] Diabetes dataset. 1990. Available at: https://www.kaggle.com/datasets/mathchi/diabetes-data-set (Accessed: 27 May 2024).

[37] Kweon S, Kim Y, Jang M, Kim Y, Kim K, Choi S, *et al*. Data resource profile: the Korea National Health and Nutrition Examination Survey (KNHANES). International Journal of Epidemiology. 2014; 43: 69–77.

[38] Domingos P. A few useful things to know about machine learning. Communications of the ACM. 2012; 55: 78–87.

[39] Bergman E, Purucker L, Hutter F. Don't waste your time: early stopping cross-validation. 2024. Available at: http://arxiv.org/abs/2405.03389 (Accessed: 18 June 2024).

[40] Géron A. Hands-on machine learning with scikit-learn, keras, and tensorflow. 2nd edn. O'Reilly Media, Inc.: Sebastopol, CA, USA. 2019.

[41] Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intelligent Data Analysis. 2002; 6: 429–449.

[42] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002; 16: 321–357.

[43] Kursa MB, Jankowski A, Rudnicki W. Boruta—a system for feature selection. Fundamenta Informaticae. 2010; 101: 271–285.

[44] Keany E. BorutaShap: a wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values. 2020. Available at: https://doi.org/10.5281/zenodo.4247618 (Accessed: 27 May 2024).

[45] Wolpert DH. Stacked generalization. Neural Networks. 1992; 5: 241–259.

[46] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery: New York, NY, USA. 2019.

[47] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing & Management. 2009; 45: 427–437.

[48] Aumann RJ, Hart S. Handbook of game theory with economic applications. Volume 4, 2021. Elsevier: Amsterdam. 1992.

[49] shap/shap. 2024. Available at: https://github.com/shap/shap (Accessed: 28 May 2024).

[50] Feller S, Boeing H, Pischon T. Body mass index, waist circumference, and the risk of type 2 diabetes mellitus. Deutsches Ärzteblatt International. 2010; 107: 470–476.

[51] Johnson RJ, Nakagawa T, Sanchez-Lozada LG, Shafiu M, Sundaram S, Le M, *et al*. Sugar, uric acid, and the etiology of diabetes and obesity. Diabetes. 2013; 62: 3307–3315.

[52] Tseng C-H. Correlation of uric acid and urinary albumin excretion rate in patients with type 2 diabetes mellitus in Taiwan. Kidney International. 2005; 68: 796–801.

[53] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery. 2019; 9: e1312.

[54] Muddamsetty SM, Jahromi MNS, Moeslund TB. Expert level evaluations for explainable AI (XAI) methods in the medical domain. Pattern Recognition. ICPR International Workshops and Challenges. 21 February 2021. Springer International Publishing: Cham. 2021.

How to cite this article: Vinh Quang Tran, Younsung Choi, Haewon Byeon. Explainable stacking ensemble with feature tokenizer transformers for men's diabetes prediction. Journal of Men's Health. 2024; 20(11): 38-56. doi: 10.22514/jomh.2024.184.