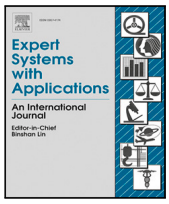
Expert Systems With Applications 274 (2025) 126899

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Dual-stage explainable ensemble learning model for diabetes diagnosis

Ibrahim A. Elgendy [a,*,1], Mohamed Hosny [a,1], Mousa Ahmad Albashrawi [a], Shrooq Alsenan [b]

[a] IRC for Finance and Digital Economy, KFUPM Business School, King Fahd University of Petroleum and Minerals, Dhahran, 31261, Saudi Arabia
[b] Information Systems Department, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, 11671, Saudi Arabia

ARTICLE INFO

Dataset link: https://physionet.org/content/mimiciv/3.1/, https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data

Keywords:
Diabetes diagnosis
Ensemble learning
Explainable artificial intelligence
Autoencoder
Healthcare

ABSTRACT

Early diagnosis of diabetes is crucial for effective management and prevention of complications. However, traditional diagnostic methods are often constrained by the complexity of clinical datasets. To this end, this study proposes a novel explainable machine learning (ML) framework to enhance diabetes prediction. Specifically, the developed methodology involves the detection of outliers using local outlier factor and data reconstruction through a sparse autoencoder. Subsequently, multiple imputation strategies are employed to effectively address missing or erroneous data, while the synthetic minority oversampling technique is applied to mitigate class imbalance. Afterward, a stacking ensemble model, consisting of seven base ML models, is developed for classification, and the outputs of these base models are aggregated using four meta models. To enhance interpretability, two layers of model explainability are implemented. Feature importance analysis is conducted to identify the significance of input variables and Shapley additive explanations is employed to assess the contribution of each base model to the meta model predictions. The results demonstrated that replacing missing data with zeros or mean values led to a noticeable decrease in accuracy compared to K-nearest neighbor imputation or removing samples. Notably, hypertension and kidney failure are pivotal features in the diabetes diagnosis process. Among the base models, Extra Trees model had the most significant impact on the meta model decisions. The stacking multi-layer perceptron model achieved the highest accuracy of 92.54% for diabetes detection, surpassing the performance of standalone ML techniques. This approach enhances diagnostic precision and provides transparency in model predictions, essential for clinical applications.

1. Introduction

Diabetes is a rapidly growing global health concern, affecting millions of individuals worldwide and posing significant challenges to public health systems. According to the World Health Organization (WHO), approximately 1.5 million people die from diabetes each year, and the number of cases continues to rise steadily (Khanam & Foo, 2021). Typically, this metabolic disorder manifests itself as high blood glucose levels caused by either insufficient insulin production (Type 1 diabetes mellitus (T1DM)) or an inability to effectively use insulin (Type 2 diabetes mellitus (T2DM)). Additionally, the long-term complications of uncontrolled diabetes include cardiovascular diseases, kidney failure and nerve damage, making early diagnosis and management crucial (Holt & Flyvbjerg, 2024).

Early detection of diabetes and appropriate intervention are significant to prevent these complications and adverse health outcomes associated with the condition. Traditional diagnostic approaches are often time-consuming and prone to errors due to their reliance on subjective evaluations and invasive procedures. Recent advances in machine learning (ML) provide promising methods for the early prediction and classification of diabetes. Utilizing these methods enables the analysis of complicated healthcare data to identify patterns that are not easily discernible through conventional diagnostic methods (Jaiswal, Negi, & Pal, 2021; Nimmagadda, Suryanarayana, Kumar, Anudeep, & Sai, 2024). With the increasing availability of medical data, including patient records and clinical test results, ML models can produce robust frameworks for improving diagnostic accuracy and assisting healthcare professionals in developing prompt interventions (Asteris et al., 2023; Gavriilaki et al., 2021). Consequently, by utilizing the capabilities of ML, researchers aim to develop automated systems that can precisely identify individuals at risk of developing diabetes (Wee, Sivakumar, Lim, Wong, & Juwono, 2024).

Although conventional ML algorithms such as artificial neural network (ANN) (Reza, Amin, Yasmin, Kulsum, & Ruhi, 2024) and support vector machines (SVM) (Nadeem et al., 2021; Yahyaoui, Jamil,

* Corresponding author.
E-mail addresses: ibrahim.elgendy@kfupm.edu.sa (I.A. Elgendy), mohamed.elnogomy@kfupm.edu.sa (M. Hosny), bishrama@kfupm.edu.sa (M.A. Albashrawi), shaalsenan@pnu.edu.sa (S. Alsenan).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.eswa.2025.126899
Received 26 December 2024; Received in revised form 31 January 2025; Accepted 12 February 2025
Available online 22 February 2025

Rasheed, & Yesiltepe, 2019) are simple and easier to interpret, the complexity of medical datasets pose significant challenges. Medical data frequently exhibit nonlinear patterns, missing values and class imbalances, all of which can impede the performance of individual ML models (Asteris et al., 2022). Therefore, they often lack the predictive power and robustness required for diabetes diagnosis. This is because diabetes prediction involves interdependent nonlinear feature interactions that single classifiers may oversimplify. Additionally, single classifiers are prone to overfitting, particularly with imbalanced and noisy data. As a result, ensemble learning models have gained attraction in diabetes diagnostics, in which they combine multiple base models to produce a stronger and more generalizable predictive model (Alnowaiser, 2024; Hasan, Alam, Das, Hossain, & Hasan, 2020). The complementary strengths of diverse base models, such as gradient boosting for handling outliers and k-nearest neighbors (KNN) for capturing local patterns, further enhance the ability to generalize across different datasets.

Despite the growing popularity of ensemble strategies, there is still room for refining existing ensemble methods to enhance performance, explainability and applicability. Many studies relied on simple voting strategies (Alnowaiser, 2024; Kalagotla, Gangashetty, & Giridhar, 2021), which fail to dynamically optimize the contributions of base models. At the same time, other studies employed advanced techniques like Bayesian optimization and feature selection methods that risk excluding clinically important features and introducing bias (Deberneh & Kim, 2021; Zhou, Xin, & Li, 2023). Additionally, most frameworks did not utilize advanced data reconstruction methods to effectively handle missing and corrupted data. Explainability is another key limitation, as existing studies focused primarily on the impact of input features without analyzing the relationship between base models and different meta models. In other words, existing studies concentrated on evaluating the overall performance of ensemble methods (Alnowaiser, 2024; Nadeem et al., 2021; Zou et al., 2018), often neglecting to explore how the base and meta models influenced the obtained predictions. To the best of our knowledge, no study has systematically provided an explainable framework to assess the contributions of individual base models to different meta models within an ensemble strategy for diabetes diagnosis. This lack of attention to explainability limits the ability to fully understand and optimize ensemble configurations, leaving a critical aspect of ensemble learning unaddressed.

Therefore, our study aims to overcome existing limitations by utilizing an explainable stacking ensemble algorithm for diabetes prediction. We introduce an ensemble strategy integrated with several meta models and explainability techniques, enhancing both accuracy and real-world applicability. In this way, adaptive boosting (AdaBoost), gradient boosting (GBoost), histogram-based gradient boosting (HistGBoost), extra trees, categorical boosting (CatBoost), extreme gradient boosting (XGBoost) and KNN were adopted as the base models within the stacking approach. The selection of these models is strategic, as they represent prominent techniques in diabetes prediction (Alnowaiser, 2024; Lai, Huang, Keshavjee, Guergachi, & Gao, 2019; Sarwar, Kamal, Hamid, & Shah, 2018; Yahyaoui et al., 2019). Furthermore, these seven models exhibit complementary strengths and weaknesses, enhancing the effectiveness of the developed stacking model. Unlike models that rely on simple voting strategies, our method employs ML techniques as the meta models. This allows the meta model to learn the optimal combination of base model predictions, leveraging their strengths while minimizing their weaknesses. By dynamically assigning weights based on the unique contributions of each base model, the proposed approach achieves superior performance and robustness compared to static voting techniques. Besides, the proposed model eliminates the dependency on traditional feature selection techniques, which can introduce bias. Instead, our approach retains all features and relies on explainability techniques to analyze their contributions. This makes the model more suitable for clinical implementation, where decisions must consider the full scope of patient data. A significant innovation in the proposed model is the integration of advanced data reconstruction techniques using sparse autoencoders. These autoencoders effectively handle data reconstruction by learning latent representations, ensuring that the reconstructed dataset captures meaningful patterns while reducing noise. Moreover, dual-stage explainability is employed to maintain transparency at both the feature and model levels, addressing the gap left by previous studies. To assess the performance of the proposed model, we compared the obtained results with those from base models and the conventional simple average ensemble (SAE) model. This comparative analysis is crucial for ensuring reliable results and evaluating our model against established methods for diabetes detection. This study provides key contributions as follows:

- Using two different real-world datasets (i.e., Medical Information Mart for Intensive Care (MIMIC)-IV and Pima Indians Diabetes (PID)) to demonstrate the model robustness across diverse feature sets. Both datasets offer challenging scenarios with missing values and unbalanced class distribution.
- Our study introduced a unique approach to handle missing and outlier data using an autoencoder-based reconstruction technique combined with four post-processing imputation strategies.
- The class imbalance problem is tackled using the synthetic minority oversampling technique (SMOTE).
- Seven longstanding ML models are initially trained independently, and their predictions are subsequently fused using four meta models (i.e., Bagging, random forest (RF), multi-layer perceptron (MLP) and logistic regression (LR)).
- The developed two-step approach facilitated the capture of higher-order interactions and non-linearities that individual models might miss. Multiple meta models are explored to ensure that the chosen architecture provides optimal accuracy in predicting diabetes outcomes.
- A novel two-layer explainability framework is designed to ensure transparency. Feature importance analysis and Shapley additive explanations (SHAP) techniques are employed to assess the contributions of input features and base models to the final diagnosis decision.
- The proposed model achieved an accuracy of 92.54% and 87.58% on the MIMIC-IV and PID datasets, respectively, demonstrating superior performance compared to existing ML techniques.

The reminder of this article is organized as follows. Section 2 provides a brief overview of the literature review in diabetes prediction. Section 3 outlines the detailed methodology used in this study. The experimental results are then analyzed and discussed in Section 4, with subsequent implications outlined in Section 5. Lastly, Section 6 concludes the main contributions and suggests potential directions for future research.

2. Literature review

Nowadays, the integration of ML has significantly transformed the healthcare sector, particularly in the domain of disease detection (Heinrichs & Eickhoff, 2020; Mainali, Darsie, & Smetana, 2021). ML techniques have facilitated the creation of accurate and dependable systems tailored to medical applications. A prominent area of focus within this field is the prediction of diabetes, where various advanced algorithms have been proposed to improve diagnostic accuracy and patient outcomes (Guan et al., 2024; Patil et al., 2024). These innovations have demonstrated the potential of ML to support clinical decision-making. For Instance, Kannadasan, Edla, and Kuppili (2019) introduced a deep neural network-based (DNN) approach for classifying T2DM data. The classification was performed using a softmax layer. Meanwhile, Talari et al. (2024) introduced an innovative classification model designed to improve the prediction accuracy of T2DM. Their model integrated the sequential minimal optimization for classification and the SMOTE for addressing class imbalance.

Table 1
Summary of the state-of-the-art methods developed for automated diabetes diagnosis.

Reference	Dataset	Size	Methodology	Accuracy	Limitations
Kannadasan et al. (2019)	PID	768	Feature selection + DNN	86.26%	Practical implementation, Limited dataset size, Model validation, Explainability.
Kalagotla et al. (2021)	PID	768	Feature selection + Voting stacked ensemble	78.2%	Limited dataset size, Practical implementation, Explainability, Low performance.
Reza et al. (2024)	PDH - PID	465 – 768	Data imputation+SMOTE+ANN-based Stacking ensemble	96.64%–77.10%	Model generalizability, Explainability.
Rubaiat, Rahman, and Hasan (2018)	PID	768	Data imputation + Feature selection + MLP	85.15%	Limited dataset size, Not robust, Model validation, Practical implementation.
Sarwar et al. (2018)	PID	768	Data imputation + Feature selection + KNN, SVM, DT, RF	77%	Limited dataset size, Not robust, Low performance.
Lai et al. (2019)	CPCSSN	13,309	Misclassification cost + LR, GBoost, RF	84.7%	Not robust, Overfitting, Explainability.
Zou et al. (2018)	Luzhou - PID	82,694 – 768	RF	80.84%–76.67%	Low performance, Not robust, Overfitting, Explainability.
Yahyaoui et al. (2019)	PID	768	RF, SVM, CNN	83.67%	Limited dataset size, Model generalizability, Explainability.
Singh and Singh (2020)	PID	768	Data imputation + Stacking ensemble	83.80%	Limited dataset size, Computational complexity, Explainability.
Alkalifah, Shaheen, Alotibi, Alsubait, and Alhakami (2025)	CGM	16,969	Data imputation + Feature selection+ANN, DT, LR, SVM	92.58%	Practical implementation, Not robust, Explainability.
Li, Peng, and Peng (2024)	BRFSS	438,693	Data imputation + Feature selection + Stacking ensemble	94.82%	Real time implementation, Explainability.
Ashisha, Mary, Kanaga, Andrew, and Eunice (2024)	BRFSS	438,693	Data imputation + Random oversampling + Boruta algorithm + RF, GBoost, Light GBoost, DT	92%	Potential biasing, Real time implementation, Explainability.
Moghaddam et al. (2024)	FACS	10,000	Data imputation + RF, GB, DT, MLP	89.7%	Model validation, Not robust, Explainability.
Jain, Tripathi, Pant, Anutariya, and Silpasuwanchai (2024)	Private Dataset	672	Data imputation + Soft voting stacked ensemble	99.4%	Model generalizability, Not robust, Explainability.
Kibria et al. (2022)	PID	768	Data imputation + SMOTE + Weighted stacked ensemble + SHAP	90%	Limited dataset size, Model validation.

The use of ensemble techniques for diabetes prediction has been increasingly gaining traction. Specifically, Kalagotla et al. (2021) exploited a filter-based correlation analysis for feature selection. Then, they developed a stacking model that integrates MLP, SVM and LR to produce the final predictions. Alnowaiser (2024) proposed another stacked ensemble framework aiming to address the prevalent issue of missing data in healthcare datasets. The framework incorporates KNN imputation technique to effectively handle missing values. XGBoost, RF and Extra Tree classifiers were integrated into a stacked ensemble voting framework as base learners. Whereas, Deberneh et al. developed ML model capable of predicting T2DM in the following year based on current-year data (Deberneh & Kim, 2021). They used analysis of variance and chi-square methods for feature selection, followed by an ensemble classification approach incorporating soft voting. Zhou et al. (2023) proposed a model based on Boruta algorithm and ensemble learning. Additionally, testing on an early diabetes risk prediction dataset confirmed the model's generalizability. It is clear that majority of existing models focused on improving performance at the expense of explainability, limiting their adoption in clinical practice. In contrast, Mohanty, Francis, Barik, Roy, and Saikia (2024) embedded SHAP into an ensemble model to enhance the accuracy of diabetes prediction. SHAP-based feature selection was shown to maintain high predictive performance while reducing computational complexity. The ensemble model demonstrated an accuracy of 86% using the top three features, compared to an accuracy of 84% when using all features. In parallel, Kibria, Nahiduzzaman, Goni, Ahsan, and Haider (2022) addressed this gap by proposing an explainable ensemble model for diabetes detection. They developed a weighted ensemble model combining RF and XGBoost with a soft voting classifier. Missing values were imputed using median values and class imbalance was mitigated using SMOTE. Table 1 comprehensively summarizes the related methods for automatic diabetes detection.

While ensemble methods have demonstrated promising results in diabetes prediction, several limitations restrict their clinical applicability and overall reliability. Many existing studies (Alnowaiser, 2024; Jain et al., 2024; Kalagotla et al., 2021) rely on simple voting strategies to combine base model predictions. However, this approach lacks the ability to optimize the contributions of each base model dynamically. Voting strategies treat all models equally or assign fixed weights, which may not consider the strengths and weaknesses of individual base models. As a result, these methods often fail to fully exploit the potential of ensemble learning for complex datasets as in our study. Other studies (Ashisha et al., 2024; Deberneh & Kim, 2021; Zhou et al., 2023) utilized Bayesian optimization and Boruta feature selection to enhance the performance. However, these approaches introduce additional complexity and potential bias. Relying on feature selection methods may inadvertently exclude clinically significant features, which can limit the trustworthiness of the model in real-world medical applications. Moreover, one notable research gap in existing diabetes prediction studies is the inconsistency in handling feature selection. Some studies have incorporated feature selection methods (Kalagotla et al., 2021; Mohanty et al., 2024; Sarwar et al., 2018; Zhou et al., 2023), identifying a subset of relevant features to improve model performance, while other studies have reported the use of all available features, arguing that every feature is important and contributes to the classification task (Zou et al., 2018). This conflicting viewpoint raises ambiguity in determining the optimal workflow for diabetes prediction models. Also, existing ensemble frameworks failed to utilize advanced data reconstruction techniques, such as autoencoders, which are crucial for handling missing and erroneous data effectively. While methods like KNN imputation and median value replacement are employed, these basic approaches may not sufficiently address the complexities of medical datasets, which often exhibit nonlinear patterns and dependencies. Furthermore, many current approaches neglect the critical role of

data reconstruction and imputation techniques (Kalagotla et al., 2021; Kannadasan et al., 2019; Yahyaoui et al., 2019).

Many of the utilized datasets in diabetes prediction studies are of limited size (Jain et al., 2024; Kalagotla et al., 2021; Kannadasan et al., 2019; Kibria et al., 2022; Rubaiat et al., 2018; Sarwar et al., 2018; Talari et al., 2024), resulting in outcomes that may not be widely applicable to the broader population. The limited dataset sizes frequently result in models that excel in particular situations but may encounter difficulties in generalization when applied to bigger patient groups. Additionally, the majority of the presented findings are derived from an individual diabetes dataset (Alkalifah et al., 2025; Alnowaiser, 2024; Lai et al., 2019; Li et al., 2024; Moghaddam et al., 2024; Yahyaoui et al., 2019). This dependence on a single dataset constrains the generalizability of the developed models, since they may not adequately consider the diversity and intricacy of actual medical diagnoses. In reality, distinct conditions necessitate models adept at managing a broad spectrum of inputs, a capability these narrowly focused models may not accomplish effectively. Also, several studies have adopted individual ML techniques for diabetes diagnosis (Ashisha et al., 2024; Sarwar et al., 2018; Yahyaoui et al., 2019; Zou et al., 2018). However, these models remain sensitive to random noise and often underperform when exposed to unseen data. Moreover, the existing models lack explainability (Alkalifah et al., 2025; Alnowaiser, 2024; Ashisha et al., 2024; Jain et al., 2024; Nadeem et al., 2021; Moghaddam et al., 2024; Zou et al., 2018). They do not provide clear insights into how predictions are made. The absence of interpretability in these models raises concerns about their applicability in clinical settings. While authors in Kibria et al. (2022), Mohanty et al. (2024) employed SHAP, their methods are primarily used to evaluate the impact of input features on the model. However, they did not analyze the relationship between the base and meta models. This gap leaves unexplored opportunities to understand and optimize how base models contribute to the final predictions. The lack of such insights diminishes the ability of these models to build trust with clinicians as the decision-making process remains opaque.

3. Methodology

This section outlines the proposed model pipeline for diabetes diagnosis, incorporating critical steps to enhance predictive accuracy. As depicted in Fig. 1, the process begins with thorough data cleaning and preprocessing, including detecting outliers and data reconstruction. Outliers are managed through various approaches, such as removal, replacing values with zeros or means and imputation using KNN. Further, to address the class imbalance, SMOTE technique is employed. After that, the dataset is partitioned into training and testing subsets. A stacking ensemble model is then applied to the preprocessed data. The model interpretability is maintained through a two-stage explainability framework, combining feature importance analysis and SHAP to provide insights into both input feature significance and model behavior. The optimal model is determined through final performance evaluations based on predefined metrics. Each phase of the pipeline will be elaborated upon in the subsequent subsections.

3.1. Dataset

In this study, the MIMIC-IV dataset was utilized to investigate diabetes diagnosis. It serves as a unique and valuable resource for diabetes research and contains health data collected through routine clinical care. This data offers a more accurate representation of real-world healthcare than contrived or simulated environments. The MIMIC-IV dataset collected information from two distinct in-hospital sources, a comprehensive custom-built EHR system encompassing the entire hospital and an EHR system specific to the intensive care unit. This multifaceted data collection approach not only captures initial admissions but also provides a broad range of patient interactions with the healthcare system, enhancing the scope of the study.

Accurate and consistent identification of diabetes patients is pivotal for both the reliability of research and effective clinical management. The International Classification of Diseases, tenth revision (ICD-10), provides a globally accepted coding framework, ensuring precise diagnoses and facilitating data comparability across various healthcare settings. Diabetes is categorized under multiple ICD-10 codes, each representing a different type or related condition. Utilizing these codes allows us to ensure that our dataset accurately captures the presence of diabetes for each individual, thereby enhancing the precision of the analysis. This standardized methodology guarantees the reliability of the results. Additionally, the use of ICD-10 codes enables the differentiation between different forms of diabetes and their related complications. The resulting dataset includes a large patient population, with 37,246 individuals, of which only 7472 have a documented diagnosis of diabetes. Specifically, the dataset contains ten key physiological indicators to predict diabetes occurrence, offering a strong basis for analyzing health metrics related to diabetes. These indicators encompass gender, age, race, systolic blood pressure, diastolic blood pressure, body mass index (BMI), hypertension, kidney failure, pregnancy and smoking status. Table 2 presents a detailed overview of the employed features.

Besides, Fig. 2 displays the correlogram depicting the linear relationship between input features and diagnosis using the Pearson correlation coefficient. This analysis aids in pinpointing the key input indicators vital for diagnosing the diabetes. The results revealed that age, kidney failure, hypertension and systolic blood pressure take the spotlight and have the highest magnitude of Pearson correlation coefficient of 0.2328, 0.2235, 0.156 and 0.1113, respectively. Therefore, these parameters are the most effective features for detecting diabetes. Moreover, gender has shown to be a valuable indicator for diabetes identification.

3.2. Preprocessing

3.2.1. Local outlier factor

Outliers can significantly affect the performance of the ML models, and thus, their detection and removal are paramount. Accordingly, the first step in the preprocessing stage is to remove outliers using the local outlier factor (LOF) algorithm. LOF identifies outliers by measuring the local density deviation of a data point relative to its neighbors. The LOF score quantifies how much a data point deviates from the local density of the nearest neighbors, with higher scores indicating outliers. Given the dataset $X = \{x_1, x_2, \ldots, x_n\}$, where each $x_i \in \mathbb{R}^m$ denotes a data sample with m features. For each data point x_i, the distance to all other points in the dataset is computed. The set $N_k(x_i)$, representing the k-nearest neighbors of x_i, is then identified. The distance between a point x_i and the neighbor x_j is denoted as $d(x_i, x_j)$, and the distance to the k^{th} nearest neighbor of x_j is represented by $d_k(x_j)$. The neighborhood set is defined as $N_k(x_i) = \{x_j : d(x_i, x_j)$ is among the k smallest distances$\}$.

The local reachability density (LRD) of a data point x_i is the inverse of the average reachability distance from x_i to neighbors and is computed as follows:

$$\mathrm{LRD}(x_i) = \left(\frac{1}{k} \sum_{x_j \in N_k(x_i)} \max(d(x_i, x_j), d_k(x_j)) \right)^{-1} \tag{1}$$

Here, $d(x_i, x_j)$ represents the distance between x_i and neighbor x_j, and $d_k(x_j)$ denotes the distance to the k^{th} nearest neighbor of x_j. The LOF score for each point x_i is then calculated by comparing the LRD of x_i with the LRD of the neighbors, given by:

$$\mathrm{LOF}(x_i) = \frac{1}{k} \sum_{x_j \in N_k(x_i)} \frac{\mathrm{LRD}(x_j)}{\mathrm{LRD}(x_i)} \tag{2}$$

To detect outliers, a threshold is set based on the LOF scores. Points with LOF scores exceeding this threshold are classified as outliers. The

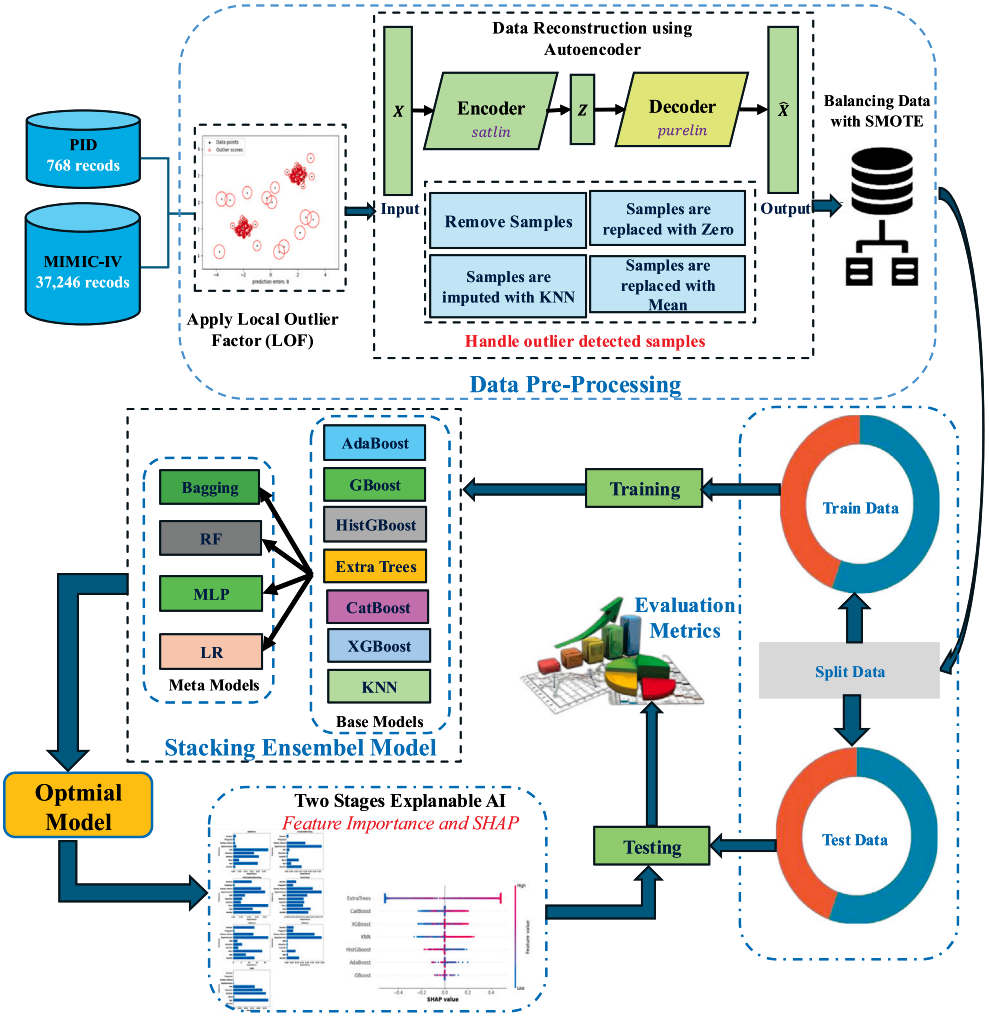


Fig. 1. Block Diagram of the proposed diabetes prediction system.

Table 2
Illustration of the MIMIC-IV diabetes dataset.

Feature	Description	Range
Gender	Binary indicator of gender (0 = Female, 1 = Male)	0–1
Age	Age at admission in years	18–101
Race	Categorical variable for race (0 = White, 1 = Black, 2 = Hispanic, 3 = Asian, 4 = Other)	0–5
Systolic	Systolic Blood Pressure (mmHg)	31–240
Diastolic	Diastolic Blood Pressure (mmHg)	0–147
BMI	Body Mass Index (weight in kg/height in m^2)	9.9–48.4
Hypertension	Indicator for hypertension status (0 = No, 1 = Yes)	0–1
Kidney Failure	Indicator for kidney failure (0 = No, 1 = Yes)	0–1
Pregnant	Indicator of pregnancy (0 = Not pregnant, 1 = Pregnant)	0–1
Smoker	Indicator of smoking status (0 = Non-smoker, 1 = Smoker)	0–1
Diagnosis	Indicator of diabetes diagnosis (0 = No diabetes, 1 = Diabetes)	0–1

cleaned dataset X_{clean} is obtained by excluding the outliers from the original dataset: $X_{\text{clean}} = \{x_i \in X : \text{LOF}(x_i) \leq \text{threshold}\}$. In this implementation, the number of neighbors k is set to 30. The final dataset, after outlier removal, is used for further analysis.

3.2.2. Autoencoder-based data reconstruction

A sparse autoencoder-based approach was applied for further data reconstruction, providing several advantages in the data preprocessing. Autoencoder is a type of unsupervised neural network which learns a compressed representation of the data to reconstruct the input from the compressed features. Autoencoder has the ability to effectively handle high-dimensional data by learning a lower-dimensional representation that captures the essential structure of the data while discarding noise and irrelevant information (Kannadasan et al., 2019). Moreover, the reconstruction error between the original data and the reconstructed data serves as a mechanism to identify outliers, allowing for more accurate data cleaning. Given the dataset $X = \{x_1, x_2, \ldots, x_n\}$, where $x_i \in \mathbb{R}^m$ is a data sample of m features, an autoencoder is trained to minimize the reconstruction error, defined as the difference between the input x_i and the reconstructed data $\hat{x}_i$. We utilize an autoencoder characterized by a hidden layer of size 10. The encoder employs the saturating linear function (satlin), while the decoder utilizes the linear function (purelin). Incorporating L2 weight regularization (0.01) and sparsity regularization (1) across 100 epochs further enhances the

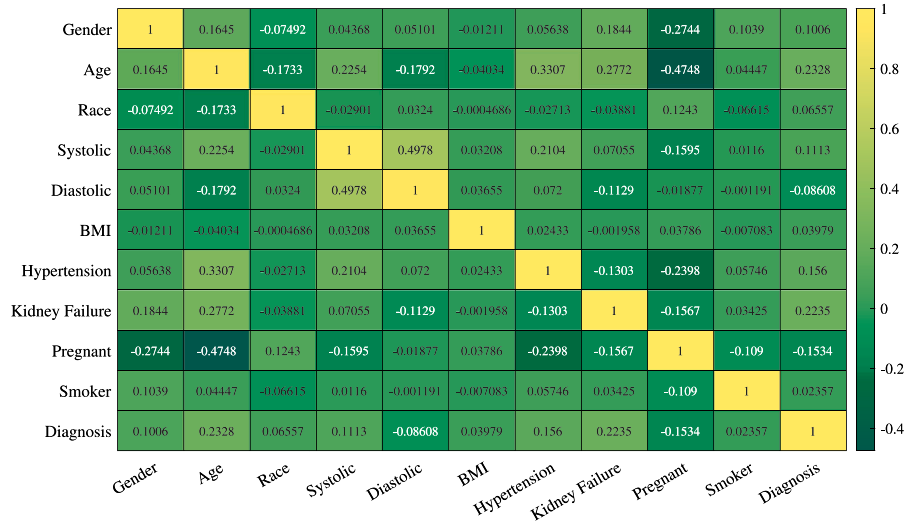


Fig. 2. The correlation matrix between the input features.

model performance by preventing overfitting and promoting meaningful feature extraction. Additionally, defining a sparsity proportion of 0.9 encourages a significant degree of sparsity within the hidden layer activations, allowing the model to focus on the most important features. The autoencoder compresses the input x_i into a hidden representation z_i using the encoder function as follows:

$$z_i = f_{\text{enc}}(x_i) = \text{satlin}(W_e x_i + b_e) \tag{3}$$

Where W_e and b_e are the weights and biases of the encoder layer. The decoder reconstructs the input $\hat{x_i}$ from the hidden representation:

$$\hat{x_i} = f_{\text{dec}}(z_i) = \text{purelin}(W_d z_i + b_d) \tag{4}$$

Where W_d and b_d are the weights and biases of the decoder layer. The reconstruction error for each data point x_i is calculated as the sum of absolute differences between the original and reconstructed values:

$$\text{Error}(x_i) = \sum_{j=1}^{m} |x_{ij} - \hat{x_{ij}}| \tag{5}$$

Data points with reconstruction errors above a certain threshold are classified as outliers. The cleaned dataset is obtained by detecting the outliers, $X_{\text{clean}} = \{x_i \in X : \text{Error}(x_i) \leq \text{threshold}\}$. After that, we employ four different techniques to handle the detected samples. The effectiveness of these techniques contributes to the overall accuracy of the diabetes diagnosis model. Firstly, data points with reconstruction errors above the threshold are directly removed. The cleaned dataset $X_{\text{clean}} = \{x_i \in X : \text{Error}(x_i) \leq \text{threshold}\}$, is then used for further analysis. In the second technique, outliers are imputed using the KNN approach (Alnowaiser, 2024). For each detected outlier x_{outlier}, we find the set of k-nearest neighbors $N_k(x_{\text{outlier}})$ from the non-outlier data points $X_{\text{non-outliers}}$. The outlier is replaced by the mean of the nearest neighbors as follows:

$$x_{\text{outlier}} = \frac{1}{k} \sum_{x_j \in N_k(x_{\text{outlier}})} x_j \tag{6}$$

In the third technique, detected outliers were replaced with the mean value of the corresponding feature column (Reza et al., 2024; Rubaiat et al., 2018; Singh & Singh, 2020). Finally, all detected outliers were replaced with zero values (Kumar, Bhusan, Singh, & kumar Choubey, 2020). For all four scenarios, the reconstructed datasets were used as input to the stacking ensemble model. The performance of the stacking model was compared across the different scenarios to assess the impact of each reconstruction technique on the model predictive accuracy.

3.2.3. Synthetic minority over-sampling

SMOTE technique is applied to address the class imbalance in the utilized dataset. SMOTE generates new instances of the minority class by interpolating between existing instances (Reza et al., 2024). The method takes the feature matrix X, a parameter N for oversampling, and k for the number of nearest neighbors as follows:

$$[X', C', X_n, C_n] = \text{smote}(X, N, k, \text{Class} = C) \tag{7}$$

Here, X' and C' represent the complete balanced dataset, while X_n and C_n denote the synthesized observations and their respective classes. Algorithm 1 illustrates the implementation of SMOTE to balance the dataset. Previous studies have shown that when SMOTE is applied without prior data preprocessing, such as handling missing values or removing outliers, often leads to the creation of erroneous synthetic data points (Wee et al., 2024). Therefore, we applied SMOTE to address the imbalance in the dataset after outlier detection and noise removal. The advantages of employing SMOTE in our preprocessing pipeline include:

1. SMOTE creates synthetic instances rather than duplicating existing samples, which introduces more variety in the minority class.
2. Reduce the risk of overfitting to the minority class in the classification algorithms, leading to improved robustness.
3. The generation of synthetic samples allows for a more representative distribution of the minority class, thereby enhancing the model's ability to learn the underlying patterns effectively while minimizing the impact of noise.
4. SMOTE provides a smoother decision boundary, which can lead to more stable predictions.

3.3. Stacking ensemble technique

Ensemble learning stands out as one of the most effective approaches for enhancing system performance. It involves the integration of diverse separate models, working in concert to ameliorate the stability and predictive prowess of the model (Kalagotla et al., 2021). Multiple algorithms are trained on the same dataset, yielding noteworthy system performance when coupled with stacking. The proposed approach elevates classification accuracy beyond the capabilities of individual algorithms (Reza et al., 2024). Specifically, our ensemble

Algorithm 1 Synthetic Minority Over-sampling Technique.

Input:
$D \leftarrow$ Original dataset (features and class labels)
$k \leftarrow$ Number of nearest neighbors
$N \leftarrow$ Number of synthetic samples to generate
Output:
$D' \leftarrow$ Augmented dataset with synthetic samples
Identify minority class samples in D
For each minority class sample x_i in D:
1. Find k nearest neighbors of x_i in the minority class.
2. For each neighbor x_j:
   (a) Generate a synthetic sample x_s using the formula:
   $$x_s = x_i + \lambda(x_j - x_i)$$
   where λ is a random value in the range [0, 1].

Add the synthetic samples x_s to the dataset D.
Return the augmented dataset D'.

Table 3
Hyperparameters of the employed ML models.

Classifier	Hyperparameters
AdaBoost	Learner: Decision Tree (Depth = 3) Estimators = 300, Learning rate = 0.5
GBoost	Estimators = 200, Learning rate = 0.5, Depth = 3, Samples leaf = 50
HistGBoost	Iterations = 300, Learning rate = 0.5, Depth = 3, Samples leaf = 30, Random state = 42
Extra Trees	Estimators = 100, Random state = 42
CatBoost	Iterations = 300, Learning rate = 0.5, Depth = 5, Random state = 42
XGBoost	Estimators = 300, Learning rate = 0.5, Depth = 5, Random state = 42
KNN	Neighbors = 3, Kernel= Cityblock
Bagging	Learner: Decision Tree(Depth = 3) Estimators = 100
RF	Estimators = 100, Depth = 5, Samples split = 10, Random state = 42
MLP	Hidden layer sizes = [100 50], Activation= Relu, Solver = Adam, Learning rate = 0.001, Iterations = 500, Random state = 42
LR	C = 100, Penalty = l2, Solver = lbfgs, Iterations = 1000, Random state = 42

approach incorporated several base ML models with diverse problem-solving abilities. The aim was to extract meta features from each model. Subsequently, these meta features were inputted into a meta model, culminating in the final classification step.

Herein, the architecture of the proposed model consisted of seven base models, namely, AdaBoost, GBoost, HistGBoost, Extra Trees, CatBoost, XGBoost, and KNN. The outputs from these models were then fed into one of four meta models, namely, Bagging, RF, MLP, and LR. Given the dataset $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{n}$, where $X_i \in \mathbb{R}^d$ is a feature vector and $y_i \in \{0, 1\}$ is the class label, the base models $\{M_1, M_2, \ldots, M_7\}$ were trained on the training data to produce predictions $\hat{y}_j^{(i)} = M_j(X_i)$ for each data point X_i. The meta model M_{meta} was trained on the predictions from the base models, $\hat{y}_{\text{meta}} = M_{\text{meta}}(\hat{y}_1^{(i)}, \hat{y}_2^{(i)}, \ldots, \hat{y}_7^{(i)})$. The meta model learned the predictions of the base models to make the final decision (Singh & Singh, 2020). Table 3 displays the implemented ML models along with their corresponding hyperparameters.

The stacking ensemble method offers several key advantages. First, it ensures a diverse set of learners that capture various aspects of the data. Additionally, the meta model within the stacking architecture learns from the mistakes made by the base models. It identifies patterns in the errors of individual base models and corrects them, thereby enhancing the overall classification accuracy. Another benefit is the flexibility of stacking, as it allows for the use of any type of base model and meta model. This adaptability facilitates better model selection and optimization tailored to the specific characteristics of the dataset. Finally, stacking reduces the risk of overfitting compared to using individual models, particularly when applied to large datasets as in our study (Asteris, Gandomi, Armaghani, Kokoris, et al., 2024; Asteris, Gandomi, Armaghani, Tsoukalas, et al., 2024). In contrast to the stacking ensemble model, the SAE model works by averaging the predictions of the seven base models. The final prediction is calculated as $\hat{y}_{\text{avg}} = \frac{1}{7}\sum_{j=1}^{7} \hat{y}_j^{(i)}$. This approach is computationally less expensive than stacking but does not capture the complex relationships between base models like a meta model does.

3.4. Explainer

In this study, we employed feature importance and SHAP to explain the impact of the input features and base model predictions on the classification outcomes, respectively (Lundberg & Lee, 2017). Both methods offer insights into how individual features contribute to the model decision-making process, but they differ in their approaches. Traditional feature importance techniques evaluate the significance of the input features in contributing to each base model prediction, providing an understanding of the underlying data dynamics. Two prominent methods were employed, permutation importance and tree-based feature importance. These techniques offer complementary perceptions into the model behavior. Permutation importance is a model-agnostic technique that assesses the importance of a feature by shuffling its values and measuring the effect on the model performance. The rationale behind this method is that if permuting a feature leads to a significant drop in the model performance, then the feature is vital for the model predictions. The procedure begins by calculating the model original prediction error. For each feature j, the values of that feature in the test dataset are shuffled while keeping all other features unchanged, and the model prediction error is recalculated. The importance of feature j is then defined as the increase in prediction error due to this permutation, $\text{Importance}(j) = \text{Error}_{\text{permuted}}(j) - \text{Error}_{\text{original}}$. Meanwhile, for tree-based models, feature importance is assessed based on how the model splits on each feature. The importance of a feature j is determined by the cumulative reduction in Gini impurity across all nodes where the feature is used for splitting. Specifically, importance of feature j is given by:

$$\text{Importance}(j) = \frac{\sum_{t \in T: j \text{ splits } t} p(t)\Delta G(t, j)}{\sum_{t \in T} p(t)\Delta G(t)} \tag{8}$$

Where $p(t)$ is the proportion of samples reaching node t, $\Delta G(t, j)$ is the decrease in Gini impurity from splitting node t and T is the set of all nodes. By combining these two approaches, we comprehensively understand the important indicators for diabetes diagnosis using different ML models. Feature importance techniques facilitate a deeper understanding of how each feature contributes to the final predictions. This is particularly crucial in ensemble methods, where the aggregation of diverse model outputs can obscure individual feature impacts. The importance of features like gender, age and BMI to the overall classification can be elucidated, allowing us to identify which attributes are driving the model predictions. Diabetes is a multifaceted disease influenced by various risk factors. Therefore, understanding these factors interactions and significance to the classification outcomes can provide valuable information for healthcare professionals. As a result, the proposed model can highlight critical features that influence diabetes risk, fostering trust and accountability.

Additionally, we employed SHAP to interpret the contribution of individual base models to the meta model predictions. Each base model is assigned an importance value that reflects its contribution to the prediction for a given instance. The SHAP value for a model i is defined

Algorithm 2 SHAP Implementation for Meta Model Predictions.

Input: Dataset X, Labels Y, Base models $M = \{M_1, M_2, \ldots, M_n\}$, Meta model F
Output: SHAP values for base models
Fit each base model M_i on training data X and Y
for each model M_i in M **do**
    Predict P_i using M_i on test set X
end for
Create transformed feature matrix $X_{\text{transformed}} = [P_1, P_2, \ldots, P_n]$
Fit the meta model F on $X_{\text{transformed}}$ and Y
Calculate average prediction $\phi_0(f)$ using F on X
for each base model M_i in M **do**
    Compute Shapley value $\phi_i(f, x)$ using:

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left(f_x(S \cup \{i\}) - f_x(S)\right)$$

end for

$$\text{Explanation}(x) = \phi_0(f) + \sum_{i=1}^{M} \phi_i(f,x)$$

as:

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left(f_x(S \cup \{i\}) - f_x(S)\right) \tag{9}$$

where N is the set of the base models, S is a subset of models excluding model i, and $|S|$ is the number of base models. The term $f_x(S)$ represents the prediction of the meta model using only the base models and $f_x(S \cup \{i\})$ represents the prediction of the meta model using the base model in S plus model i. The factorial terms $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ represent the number of possible sequences in which the base models can be ordered, without model i. SHAP explanations are additive, meaning that the sum of SHAP values for all base models corresponds to the difference between the prediction for an instance and the average prediction for the dataset. This can be expressed as follows:

$$\text{Explanation}(x) = \phi_0(f) + \sum_{i=1}^{M} \phi_i(f,x) \tag{10}$$

Where $\phi_0(f)$ is the average prediction over the dataset and $\sum_{i=1}^{M} \phi_i(f,x)$ represents the SHAP values for all models. Algorithm 2 illustrates the SHAP implementation in our study.

Clearly, explainability techniques offer significant advantages in enhancing model interpretability (Hosny, Elshenhab, & Maged, 2025). By integrating explainability into our analysis, we aid in better-informed clinical decisions and personalized healthcare interventions.

3.5. Model implementation and evaluation

In this study, we utilized 80/20 split for dataset partitioning, where 80% of the data was allocated for training the proposed methods, while the remaining 20% was reserved for testing. This technique is selected to align with standard practices in diabetes diagnosis research to maintain comparability with prior studies that adopted this ratio as a benchmark for evaluating the performance of their methods (Alnowaiser, 2024; Farnoodian, Moridani, & Mokhber, 2024; Mohanty et al., 2024; Reza et al., 2024; Tan, Chen, Zhang, Tang, & Liu, 2022). Also, this technique is particularly suitable for our study due to the size and diversity of the MIMIC-IV dataset, which contains data from 37,246 unique patients. This minimizes the risk of bias and guarantees a robust assessment of the proposed methods. The random split also accounts for the inherent variability in patient data, allowing the model to learn general patterns in the training set while being independently tested on unseen data.

Grid search technique was employed to optimize the hyperparameters for all the proposed techniques. Grid search is an exhaustive search method that systematically evaluates a predefined range of hyperparameters for each model, aiming to identify the combination that yields the best performance. This approach was applied uniformly across all base models and meta models in the stacking ensemble, ensuring that each model was fine-tuned to achieve the highest predictive accuracy for diabetes detection. Moreover, the performance of the proposed methods was assessed using several evaluation metrics, including accuracy, precision, sensitivity (Sens), specificity (Spec), F1 Score and area under the curve (AUC). Experiments were conducted on a device with Intel Core i7 (2.20 GHz) CPU and 8 GB of RAM using the Jupyter Notebook.

4. Results and discussion

In this section, we first evaluate the performance of the developed stacking ensemble model against the seven base ML models. We then explore the performance of four meta models within the stacking framework to gauge which combination of base and meta models yields the best results for diabetes diagnosis. Additionally, we compare the results of the proposed model across different autoencoder-based data reconstruction techniques, including removing the detected samples, replacing them with the KNN imputation technique, replacing them with the mean of the respective feature, and replacing the values with zeros. Furthermore, the proposed stacking ensemble model was compared to SAE technique. Moreover, we apply this approach to the PID dataset to validate the effectiveness and generalizability of the proposed methodology. Finally, we introduced a two-layer explainability method using feature importance and SHAP techniques to provide an understanding of the model decision-making.

Table 4 displays classification measures of the developed models using the first technique, where samples are removed. Among the base models, ExtraTrees has emerged as the top performer with accuracy of 90.38%, F1-score of 90.01% and AUC of 90.36%. This model demonstrated strong precision of 93.12%. On the other hand, although KNN exhibited the lowest accuracy of 86.36%, it obtained sensitivity of 96.54% by correctly identifying non-diabetic cases. Fig. 3 shows a comparison between the proposed methods in terms of accuracy, precision and F1-score. The proposed stacking ensemble models performed notably better than individual ML models. Stacking MLP achieved the highest overall performance with accuracy of 92.54%, F1-score of 92.29% and AUC of 92.52%. Notably, this model showed a strong balance between sensitivity (89.79%) and specificity (95.26%), indicating robustness in handling the dataset. The comparison of meta models within the stacking ensemble showed that stacking LR performed competitively with an accuracy of 92.39% and the highest sensitivity of 91.14%. These results suggest that using stacking with different meta models improves classification accuracy, particularly when outliers are removed.

Fig. 4 displays the comparison results between the stacking and SAE strategies. The proposed models consistently outperformed SAE. This signifies that the stacking ensemble approach is more effective at capturing patterns in the data than the averaging technique used in SAE. This shows that the stacking ensemble models better handle the trade-off between precision and recall. The stacking strategy uses meta models to weight the outputs of base models, while SAE simply averages predictions without learning an optimal combination. This adaptability allows the stacking models to exploit the strengths of each base model more effectively.

Table 5 provides classification measures of the proposed models using the second technique, where data samples were replaced using KNN imputation. Among the base models, ExtraTrees achieved the highest performance with an accuracy of 87.67%. ExtraTrees displayed a balanced precision of 89.66% and sensitivity of 85.20%, making it

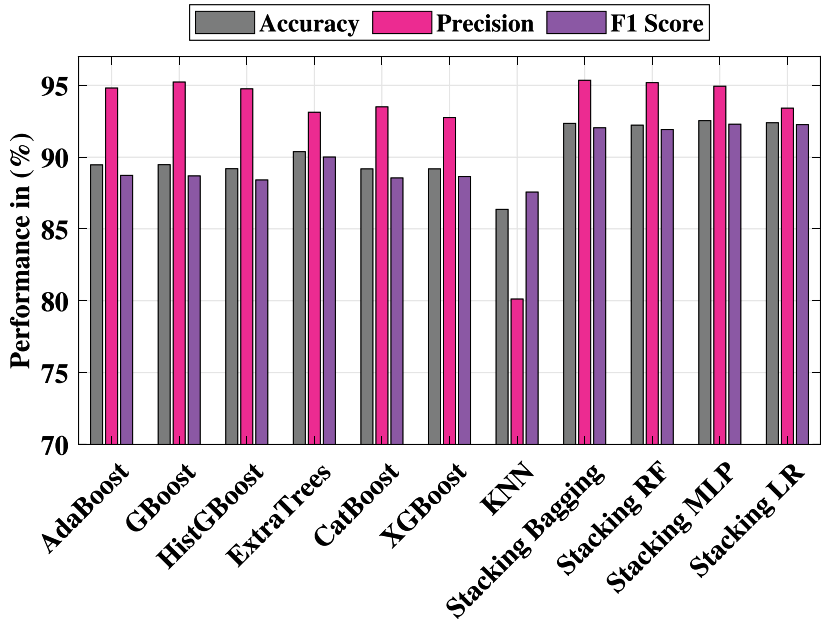


Fig. 3. Comparison between the proposed methods in terms of accuracy, precision and F1-score.

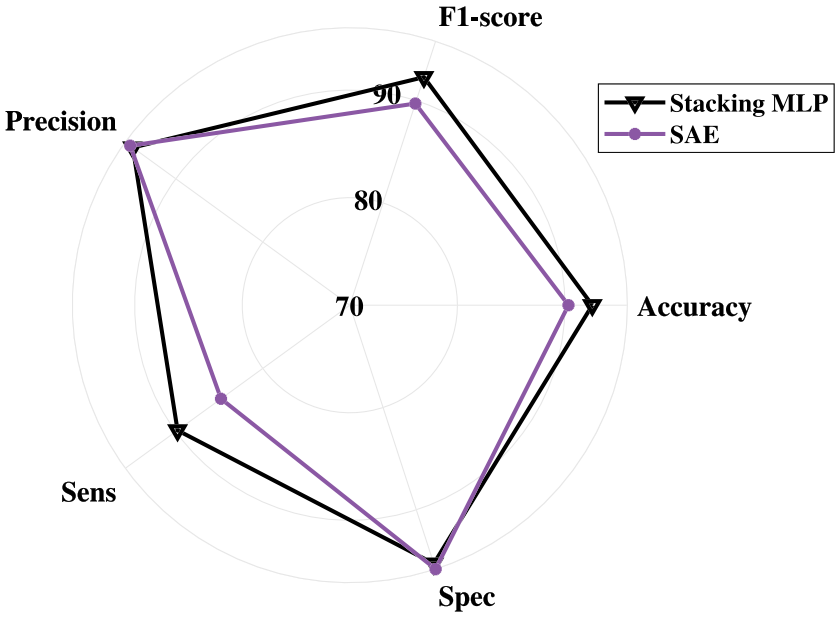


Fig. 4. Comparison results between the proposed stacking ensemble and SAE models.

one of the most reliable classifiers for diabetes identification. CatBoost followed closely with an accuracy of 86.29%. KNN showed the lowest accuracy of 82.22%. While, Stacking MLP emerged as the best classifier with the highest accuracy of 89.40%, showing robustness in handling the imputed data. Overall, the results suggest the versatility of the stacking approach in dealing with diabetes data after using KNN imputation technique.

Tables 6 and Table 7 provide the results for the last two autoencoder-based data imputation methods, where data samples were replaced by the mean and zero values, respectively. In Table 6, the stacking models consistently outperformed the traditional ML models. ExtraTrees achieved the highest accuracy of 84.09%, followed by CatBoost and XGBoost with accuracies of 83.29% and 82.72%, respectively. KNN recorded the lowest accuracy of 79.86%, with a notable drop in specificity. The stacking models exhibited improved performance across all metrics, with Stacking MLP achieving an accuracy of 85.34%. In Table 7, the results are slightly better than those from mean imputation. ExtraTrees and CatBoost maintained similar performance with accuracies of 83.96% and 83.13%, respectively. Among the stacking models, stacking LR emerged as the top performer, achieving an accuracy of 85.31%. Again, the stacking models outperformed the well-established ML classifiers using this imputation method.

Fig. 5 shows the accuracy of the implemented classifiers using the autoencoder-based techniques. It is evident that the stacking ensemble models consistently outperformed other ML models. The performance of base models like ExtraTrees, CatBoost and XGBoost remained relatively stable across different scenarios. Stacking MLP model demonstrated high precision, making it a strong candidate for accurate diabetes diagnosis using these imputation strategies. It is clear that replacing missing values with zero or the mean generally led to a

Table 4
Classification measures of the implemented models after removing data samples.

Classifier	Accuracy	F1-score	AUC	Precision	Sens	Spec
AdaBoost	89.46	88.73	89.43	94.81	83.38	95.48
GBoost	89.47	88.69	89.44	95.23	83.00	95.88
HistGBoost	89.19	88.41	89.16	94.75	82.86	95.46
ExtraTrees	90.38	90.01	90.36	93.12	87.09	93.63
CatBoost	89.18	88.55	89.16	93.50	84.10	94.21
XGBoost	89.18	88.65	89.16	92.75	84.89	93.43
KNN	86.36	87.57	86.41	80.12	96.54	76.29
SAE	90.32	89.71	90.30	95.23	84.80	95.79
Stacking Bagging	92.35	92.04	92.33	95.35	88.96	95.70
Stacking RF	92.23	91.92	92.21	95.18	88.87	95.55
Stacking MLP	92.54	92.29	92.52	94.94	89.79	95.26
Stacking LR	92.39	92.26	92.39	93.41	91.14	93.63

Table 5
Classification measures of the implemented models after using KNN imputation.

Classifier	Accuracy	F1-score	AUC	Precision	Sens	Spec
AdaBoost	85.91	84.87	85.91	91.74	78.96	92.87
GBoost	85.76	84.75	85.76	91.33	79.05	92.48
HistGBoost	85.32	84.24	85.33	91.05	78.38	92.27
ExtraTrees	87.67	87.37	87.67	89.66	85.20	90.15
CatBoost	86.29	85.43	86.29	91.24	80.31	92.27
XGBoost	85.82	85.16	85.83	89.47	81.24	90.41
KNN	82.22	83.94	82.21	76.62	92.81	71.61
SAE	87.02	86.25	87.03	91.79	81.35	92.70
Stacking Bagging	88.18	87.04	88.19	96.47	79.29	97.10
Stacking RF	88.84	88.14	88.85	94.21	82.80	94.90
Stacking MLP	89.40	89.32	89.40	90.13	88.52	90.28
Stacking LR	89.00	88.71	89.00	91.21	86.35	91.66

Table 6
Classification measures of the implemented models after replacing samples with mean values.

Classifier	Accuracy	F1-score	AUC	Precision	Sens	Spec
AdaBoost	83.36	83.83	83.35	81.62	86.16	80.55
GBoost	83.31	83.78	83.31	81.58	86.11	80.51
HistGBoost	82.70	83.06	82.70	81.45	84.73	80.66
ExtraTrees	84.09	84.80	84.09	81.26	88.67	79.50
CatBoost	83.29	83.88	83.29	81.12	86.83	79.75
XGBoost	82.72	83.46	82.71	80.10	87.11	78.31
KNN	79.86	82.55	79.85	72.89	95.17	64.52
SAE	84.00	84.48	84.00	82.09	87.02	80.98
Stacking Bagging	84.60	85.13	84.60	82.38	88.08	81.12
Stacking RF	84.95	85.48	84.95	82.64	88.52	81.37
Stacking MLP	85.34	86.08	85.33	82.02	90.56	80.10
Stacking LR	85.25	86.08	85.25	81.60	91.08	79.41

Table 7
Classification measures of the implemented models after replacing samples with zeros.

Classifier	Accuracy	F1-score	AUC	Precision	Sens	Spec
AdaBoost	83.69	84.17	83.69	81.85	86.63	80.75
LightGBoost	83.33	83.79	83.33	81.64	86.05	80.60
GBoost	82.79	83.17	82.79	81.45	84.97	80.60
ExtraTrees	83.96	84.66	83.96	81.23	88.39	79.52
CatBoost	83.13	83.70	83.13	81.07	86.52	79.75
XGBoost	83.00	83.76	83.00	80.29	87.54	78.46
KNN	79.85	82.55	79.84	72.88	95.17	64.50
SAE	83.94	84.42	83.94	82.07	86.91	80.98
Stacking Bagging	84.67	85.13	84.67	82.75	87.65	81.68
Stacking RF	84.92	85.51	84.92	82.38	88.89	80.94
Stacking MLP	85.08	85.72	85.07	82.26	89.49	80.66
Stacking LR	85.31	86.02	85.31	82.16	90.25	80.36

decrease in accuracy across the models compared to KNN imputation technique. However, the stacking models demonstrated resilience to these imputation methods, further validating their effectiveness.

To further validate the effectiveness of the proposed methodology, we used the PID dataset (*Pima Indians Diabetes Database*, 2018). This dataset is widely utilized by researchers in various studies (Kalagotla et al., 2021; Kannadasan et al., 2019; Rubaiat et al., 2018; Sarwar et al., 2018; Singh & Singh, 2020), making it a paramount benchmark for comparison and evaluation. The extensive use of this dataset across different experiments ensures that our findings can be effectively contextualized within the broader field of diabetes research. Besides, the PID dataset presents a challenging scenario due to the presence of missing values, such as zero entries for features like BMI and skin thickness. This aspect aligns with the complexity of our data, where handling missing values and outliers is also crucial. Furthermore, the PID dataset shares a common issue with our data, as it contains significantly more non-diabetic than diabetic cases. This mirrors the distribution challenges in our data and enables us to evaluate the performance of our methodology in another unbalanced data scenario. Finally, despite these similarities, the PID dataset differs from ours in that it contains unique input features, such as glucose levels, insulin and skin thickness. This variation allows us to test the robustness of our model across different feature sets and reveals the model's ability to generalize beyond the specific characteristics of the MIMIC-IV dataset. Table 8 highlights the performance of the proposed system embedded with different meta models on the PID dataset. Stacking RF model achieved the highest accuracy of 87.58%, F1-score of 86.13% and AUC of 88.00%. These results validate the strength of the RF meta model in handling the complex patterns in the PID dataset. Stacking Bagging model followed closely with an accuracy of 86.93%. Despite the lower performance in certain metrics, the overall results still validate the stacking ensemble approach with the four meta models. These results imply that the use of autoencoder for data reconstruction is highly effective for diabetes diagnosis when applied to different datasets. Furthermore, the four stacking models performed competitively, providing strong evidence that the methodology is adaptable. The high sensitivity scores across all stacking models confirm that the proposed approach is particularly adept at detecting non-diabetic cases, making it a reliable method for medical diagnostics in real-world healthcare settings.

Table 8
Performance measures of the proposed models on the PID dataset.

Classifier	Accuracy	F1-score	AUC	Precision	Sens	Spec
Stacking Bagging	86.93	85.51	87.43	80.82	90.77	84.09
Stacking RF	87.58	86.13	88.00	81.94	90.77	85.23
Stacking MLP	86.27	84.89	86.86	79.73	90.77	82.95
Stacking LR	85.62	84.51	86.49	77.92	92.31	80.68

In this study, we implemented a dual-stage explainable ensemble model. The first stage focused on using the feature importance technique to explain the significance of the input features for each base model. This step improved transparency on how various input features can influence the decision-making process of each base model. In the second stage, SHAP was applied to investigate the impact of each base model on the overall performance of the meta model. This provides a comprehensive explanation of the entire process, allowing for a deeper understanding of both feature contributions and model-level interactions within the ensemble. Fig. 6 shows the importance of the input features across the seven base ML models. Each model assigns varying degrees of importance to the different input features. By comparing these importance scores, we can better understand how the models arrive at their classification decisions.

ExtraTrees achieved the highest overall performance due to the fact that it is the only base model that gives considerable importance to all input features, capturing a broader scope of information from the dataset. This comprehensive feature utilization likely contributed to higher performance, as it successfully balanced between identifying diabetic and non-diabetic patients with sensitivity of 87.09% and specificity of 93.63%. KNN stands out for achieving the highest Sensitivity of 96.54%. KNN excels at identifying non-diabetic individuals by focusing on age and blood pressure-related features. Notably, KNN gives less importance to hypertension and kidney failure, which suggests that

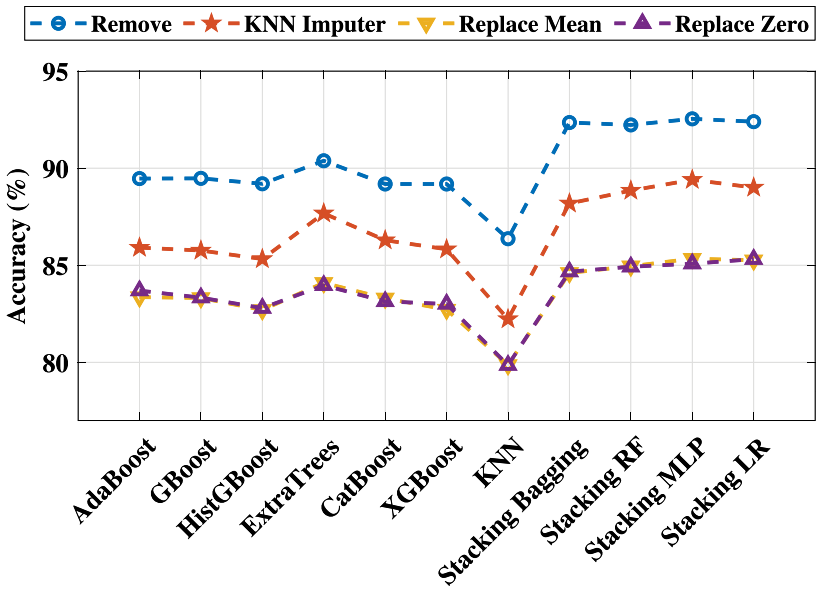


Fig. 5. Accuracy of the proposed models using different Autoencoder-based techniques.

it focuses more on general health markers to identify the absence of diabetes. GBoost showed a high specificity of 95.88%, making it very effective at identifying diabetic patients. It can be seen that GBoost assigned high importance to hypertension, kidney failure, race and gender, which are critical for identifying diabetes. These results suggest that focusing on these particular risk factors aids in identifying diabetic patients. In parallel, CatBoost focused on hypertension and kidney failure, highlighting these as strong indicators of diabetes, though it still values other features. While AdaBoost performed strongly across the board, particularly focusing on BMI, age and systolic features. HistGBoost follows a similar trend, giving weight to hypertension, race and age. XGBoost also highlights Hypertension as the most important feature. Hypertension is consistently ranked highly across most models, confirming its relevance in diabetes prediction. It is clear that, all features contribute valuable information to the classification process, with some being more critical than others depending on the specific model. Importantly, base models focus on hypertension and kidney failure for identifying diabetes. Moreover, feature importance analysis underscores the necessity of considering all available information for optimal performance in diabetes classification.

Additionally, we exploited SHAP analysis to reveal insightful relationships between the meta and base models as shown in Fig. 7. The experimental findings divulge that ExtraTrees stands out as the best individual base model. In parallel, the SHAP values demonstrate that ExtraTrees exerts the most substantial influence on the meta model predictions. This reveals the meta model ability to intelligently capitalize on the strengths of the base models. Interestingly, despite KNN relatively lower overall performance, MLP still assigns considerable importance to it. This focus is likely due to the KNN exceptional sensitivity. This indicates that the meta model strategically takes advantage of the KNN ability to accurately identify true positives. Accordingly, meta-ML models have the adaptive capability to assign sufficient weight to each base model, which enhances the overall prediction outcomes. Moreover, although AdaBoost, GBoost, and HistGBoost performed comparably well, they have relatively less influence on the final MLP predictions than XGBoost or KNN. This signifies that MLP selectively prioritizes models based on more nuanced characteristics than just general accuracy. This level of explainability is pivotal to verify the strategic selection of base models in our approach. Also, different models contribute unique strengths to the final meta model. Besides, unlike earlier studies that have evaluated base models solely based on their individual accuracy, our SHAP analysis has shown that individual metrics do not always correlate with the importance of the models in the ensemble. This underlines the importance of explainability in refining and validating model architecture choices.

Fig. 8 shows the dependency plot to highlight the interaction between base models. ExtraTrees dominates the interactions as indicated by the higher SHAP values across a broad range of other base models values. This suggests that ExtraTrees consistently provides robust and decisive information. In Fig. 8(b), the SHAP values are mostly centered around zero, indicating that ExtraTrees and GBoost contribute moderately to predictions. However, ExtraTrees shows more variability and substantial impact on certain predictions, as evidenced by the wider spread of blue points. While, in Fig. 8(c), the SHAP values for ExtraTrees are clustered above zero, implying a more positive influence on the meta model predictions. In contrast, the HistGBoost model shows more negative SHAP values, indicating a tendency to decrease prediction outputs. ExtraTrees again demonstrates a wider spread and higher influence in the positive SHAP value range. In Fig. 8(d), ExtraTrees shows a clear and consistent positive impact on the meta model predictions, with a nearly linear relationship between the SHAP and model output values. While ExtraTrees takes the lead, XGBoost complements it with valuable predictive input, especially in higher SHAP value ranges as shown in Fig. 8(f). Although XGBoost performance is slightly below ExtraTrees, the two models interact in a way that enhances prediction certainty. Unlike the smooth transition of SHAP values seen in other base models, KNN produces more discrete clusters in the SHAP distribution as shown in Fig. 8(g). KNN operates based on neighborhood data points and categorical assignments, which leads to more pronounced jumps in SHAP values rather than continuous fluctuations. This pattern aligns with KNN nature, where predictions may change drastically as input values cross classification boundaries. KNN achieved strong sensitivity, making it influential in detecting non-diabetic cases. Interaction between KNN and ExtraTrees likely reflects the MLP reliance on KNN for scenarios where sensitivity is key. These SHAP results challenge the traditional practice of focusing solely on accuracy when building stacking ensembles. Moreover, this analysis highlights the value of explainable models, particularly in fields where accuracy and sensitivity hold equal importance, such as healthcare and medical diagnostics.

Overfitting is a pivotal challenge in ML, particularly in medical diagnostics, where models may memorize patterns specific to the training data rather than generalizing to unseen data (Asteris, Skentou,

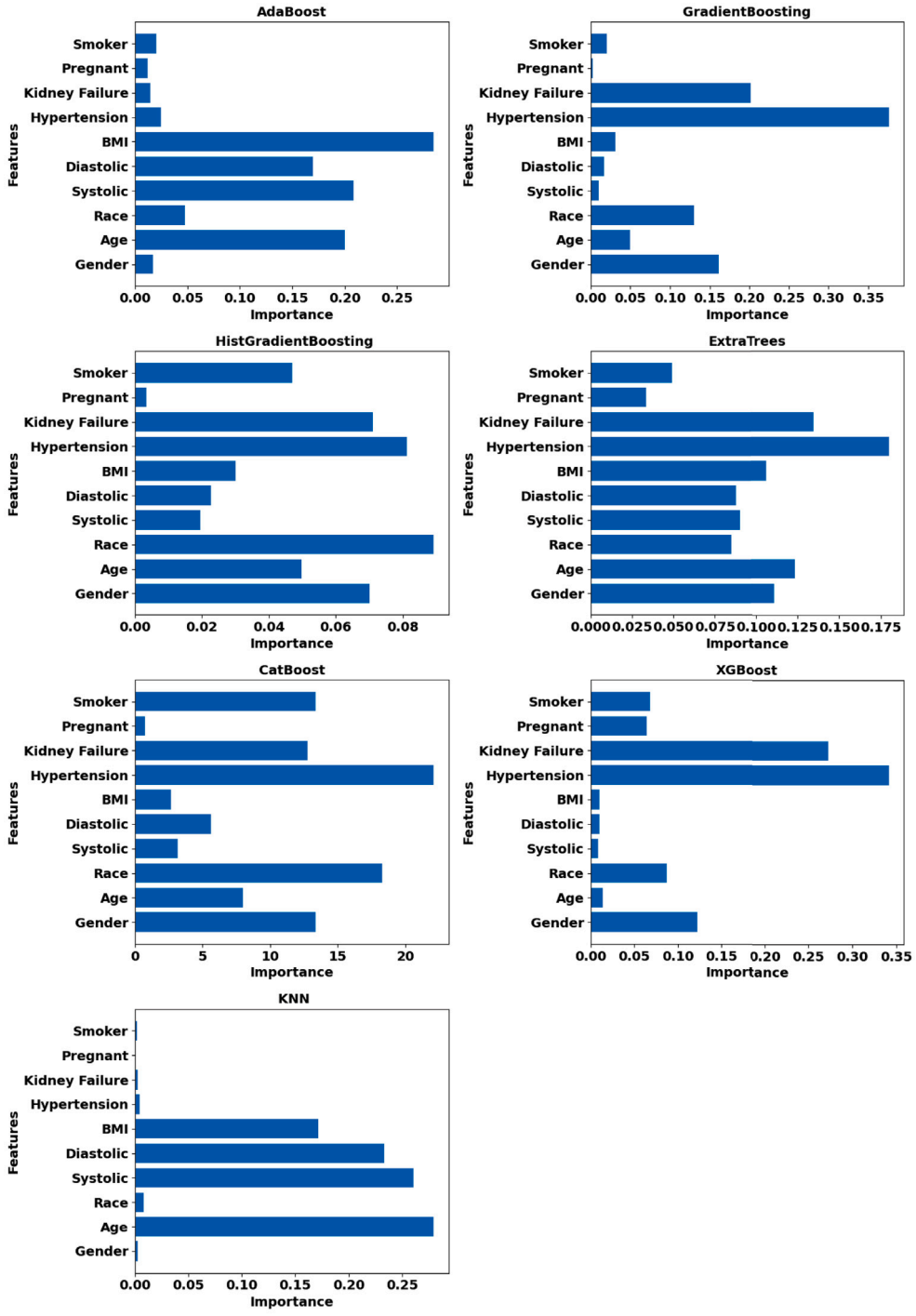


Fig. 6. Feature importance for each base model.

Bardhan, Samui, & Pilakoutas, 2021; Benzaamia, Ghrici, Rebouh, Zygouris, & Asteris, 2024). To address the potential risk of overfitting in the proposed model, several strategies were implemented to ensure unbiased performance. The independence of the training and testing datasets was strictly maintained. Each row in the dataset corresponds to a unique patient to eliminate the risk of information leakage. Besides, various data imputation techniques were carefully evaluated to address potential biases caused by missing values and further enhance the model reliability. Additionally, our methodology effectively tackled the common issue of unbalanced diabetes datasets using SMOTE. Furthermore, the designed approach inherently mitigates overfitting by combining the predictions of multiple base models through different meta models to generalize effectively. We also employed a validation framework by systematically comparing multiple meta models to identify the most effective configuration. Moreover, the external dataset was utilized to confirm the model's effectiveness. Finally, a novel two-layer explainability framework was introduced. This explainability guaranteed transparency and fostered trust in the model predictions. This comprehensive approach ensures the model's applicability across

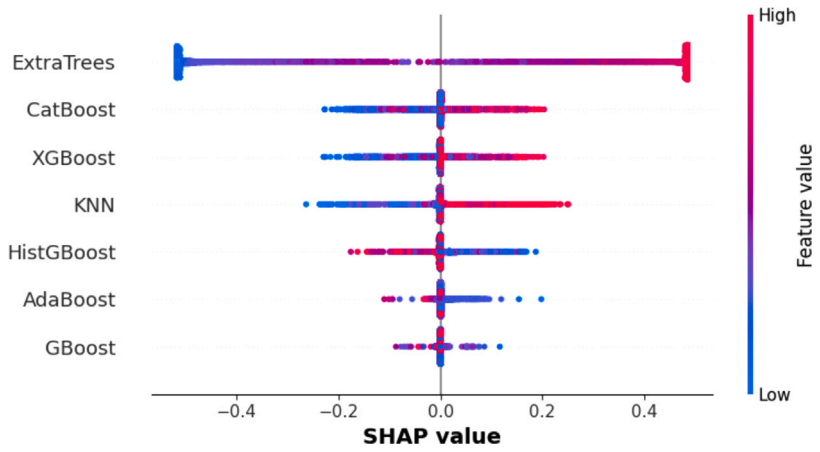


Fig. 7. Explainability of the MLP classifier using base model predictions.

diverse patient populations.

5. Implications

The integration of advanced outlier detection and data reconstruction techniques addresses a significant challenge faced by traditional diagnostic methods such as the presence of noisy and incomplete clinical data. By efficiently handling outliers and employing various imputation strategies, this study offers a more robust framework for managing missing and erroneous data to ensure that ML models can operate effectively in real-world medical settings. Moreover, the implementation of SMOTE technique to balance the dataset addresses a common issue in medical datasets. This ensures that the model performs well even when faced with a higher proportion of non-diabetic samples. Furthermore, this multi-model approach captures various underlying patterns within the clinical data, enabling the system to generalize better across patient populations. The high accuracy achieved by the proposed models demonstrates that ensemble learning frameworks hold considerable promise for diabetes diagnosis. This could translate to earlier detection of diabetes in clinical environments, potentially improving patient outcomes through timely interventions. Additionally, the system not only predicts diabetes with high accuracy but also elucidates the factors influencing these decisions through explainability techniques. This transparency is essential for clinical acceptance, as healthcare professionals need to understand how ML model reaches its conclusions before relying on it for diagnoses and treatment plans. The ability to interpret the model reasoning offers clinicians valuable perceptions while retaining their control over final diagnoses. The strategies developed in this study could be adapted for other diseases or medical conditions, further expanding the impact of this work.

The computational efficiency of the proposed model along with adaptability to various data scenarios, makes it well-suited for integration into existing clinical workflows. Besides, the model resilience in handling imputed and missing data across different datasets suggests deploying in various healthcare settings. This is particularly important in resource-limited environments where high-quality data may not always be available. Additionally, the system potential for integration into telemedicine platforms represents a transformative advancement in remote healthcare. Clinicians could remotely analyze patient data, while receiving accurate and explainable diagnostic assistance. This could greatly benefit underserved regions, ensuring that patients receive timely and accurate care regardless of location. The model cloud-based potential further enables real-time collaboration between healthcare providers and specialists, offering opportunities for rapid second opinions, especially in urgent or emergency scenarios.

However, there are some limitations to the proposed methodology. The proposed stacking ensemble approach is indeed more complex than single classifiers. However, we enhanced the model interpretability through the integration of a dual-stage explainability framework. This layered explainability ensured that the decision-making process remains transparent. Additionally, the proposed model is still less complex than deep learning models proposed in recent studies, which often involve intricate architectures (Chen, Wu, & Chiu, 2024). Thus, the proposed approach strikes a balance between performance and interpretability, making it suitable for practical applications in healthcare. Besides, although SHAP is a powerful tool for interpreting ML models, it does not establish causation between input features and the diabetes diagnosis but rather provide an attribution of feature importance based on changes in model predictions. Additionally, SHAP can be computationally expensive, especially for complex models. Furthermore, SHAP explains both individual and global behaviors which may overlook important interactions specific to individual instances (Tanim et al., 2025). Despite these drawbacks, SHAP remains valuable when used alongside other explainability methods to ensure a more holistic understanding of the model decisions. Therefore, we paired SHAP with feature importance analysis to provide comprehensive insights into the stacking model behavior.

6. Conclusion

A dual-stage explainable stacking ensemble model for diabetes diagnosis is introduced, achieving paramount advancements in both predictive performance and interpretability. Our methodology began with a comprehensive preprocessing phase, where outliers were detected using the LOF technique. Following this, autoencoder-based technique is utilized for data reconstruction. To manage anomalous data and enhance reliability, multiple imputation strategies were applied, including removal, KNN imputation and replacement of values with zeros or means. To address the class imbalance, the SMOTE technique was employed. We then developed a stacking ensemble model that integrated seven base classifiers, AdaBoost, GBoost, HistGBoost, Extra Trees, CatBoost, XGBoost and KNN, along with four meta models, Bagging, RF, MLP and Logistic Regression. The proposed two-step stacking ensemble approach helped to capture higher-order interactions and nonlinear relationships often missed by single classifiers. The model was evaluated on two significant real-world datasets to demonstrate the robustness across diverse feature sets. To ameliorate transparency, we implemented a two-layer explainability framework using feature importance and SHAP values, providing key understanding of the contributions of both input features and base models predictions. Furthermore, the study systematically compared the four meta models within the stacking framework,

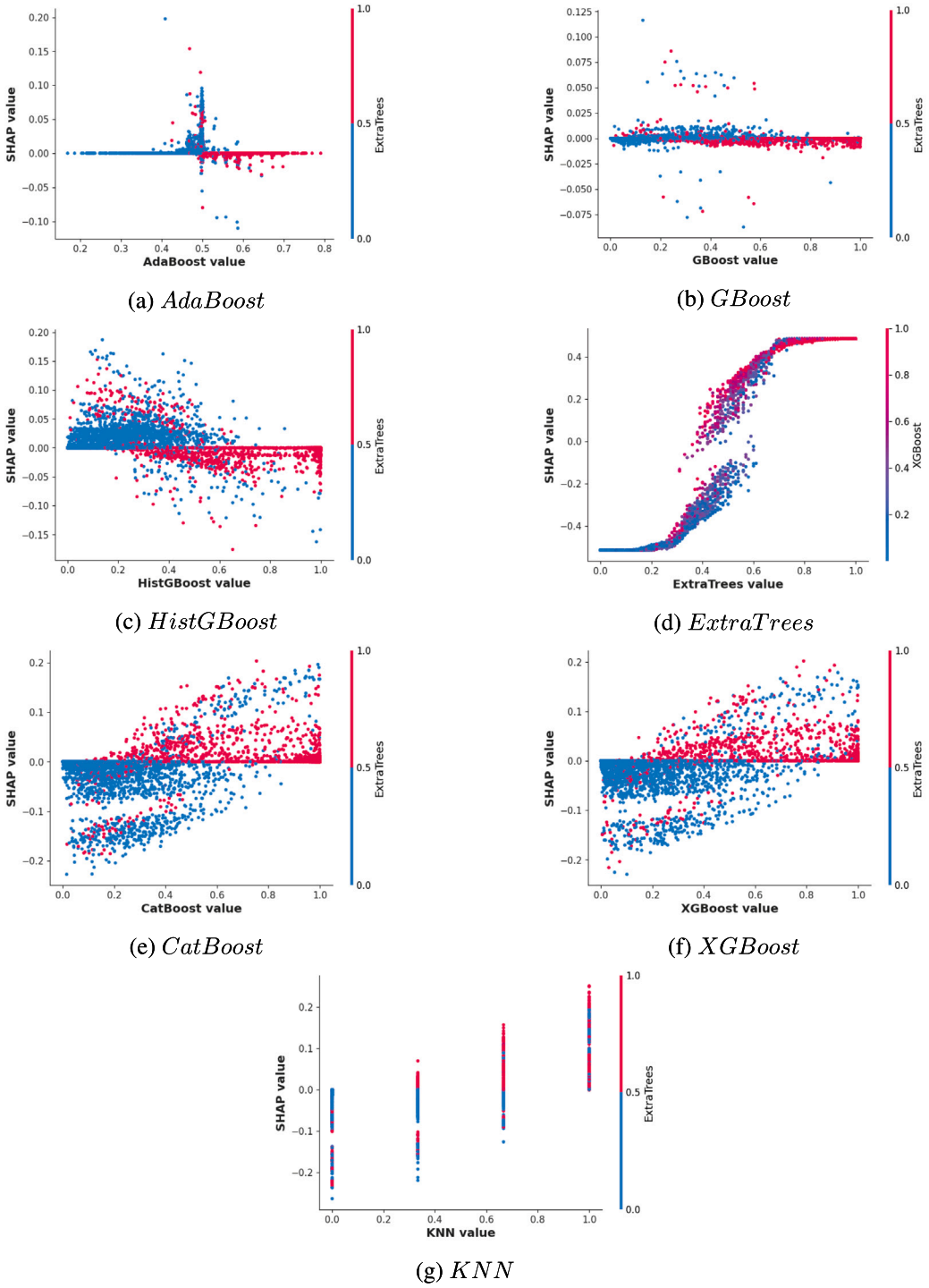


Fig. 8. Base ML models dependencies plots.

identifying the most effective combination to enhance predictive accuracy. The Stacking MLP model demonstrated the highest performance, achieving 92.54% accuracy on the MIMIC-IV dataset and 87.58% accuracy on the PID dataset. The proposed model demonstrated superior performance compared to traditional ML methods. Moreover, the stacking ensemble approach outperformed SAE by effectively exploiting a meta model to optimize the weighting of base model outputs. Besides, the experimental results demonstrated that the choice of data imputation method significantly impacts the model performance. Importantly, the experimental results revealed that hypertension and kidney failure are playing critical roles in the diabetes detection process. The SHAP-based explainability framework is pivotal in validating the strategic selection of the base models. Among the base models, ExtraTrees exerted the most substantial influence on the meta model predictions. These findings signify the importance of incorporating explainability techniques to refine and validate the model architecture. The proposed system is well-suited for integration into clinical practice by facilitating diabetes prediction and decision making.

In future work, we aim to evaluate the model on a wider variety of datasets and seek validation to ascertain the model therapeutic efficacy. While the proposed system is designed to be highly interpretable, it

could further benefit from user-friendly interfaces and additional tools to non-expert users. Finally, we plan to explore the creation of more simplified and safe iterations of this framework to facilitate adoption in smartphone-based applications.

CRediT authorship contribution statement

Ibrahim A. Elgendy: Conceptualization, Methodology, Validation, Software, Writing – original draft. **Mohamed Hosny:** Conceptualization, Methodology, Formal analysis, Software, Writing – original draft. **Mousa Ahmad Albashrawi:** Validation, Investigation, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Shrooq Alsenan:** Software, Formal analysis, Writing – review & editing.

Code availability

The source code used in this study will be available upon reasonable request from the corresponding author.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

All authors have read and agreed to the submitted version of the manuscript. This study is conducted as part of the IRC for Finance and Digital Economy, under the project number INFE2307.

Data availability

The MIMIC-IV dataset contains confidential and sensitive information, including patient data, and is therefore not publicly available. Access to the data is restricted to ensure compliance with privacy regulations and ethical standards. The datasets used in this study can be downloaded from:

https://physionet.org/content/mimiciv/3.1/ and

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data.

References

Alkalifah, B., Shaheen, M. T., Alotibi, J., Alsubait, T., & Alhakami, H. (2025). Evaluation of machine learning-based regression techniques for prediction of diabetes levels fluctuations. *Heliyon*, *11*(1).

Alnowaiser, K. (2024). Improving healthcare prediction of diabetic patients using KNN imputed features and tri-ensemble model. *IEEE Access*.

Ashisha, G., Mary, X. A., Kanaga, E. G. M., Andrew, J., & Eunice, R. J. (2024). Random oversampling-based diabetes classification via machine learning algorithms. *International Journal of Computational Intelligence Systems*, *17*(1), 270.

Asteris, P. G., Gandomi, A. H., Armaghani, D. J., Kokoris, S., Papandreadi, A. T., Roumelioti, A., et al. (2024). Prognosis of COVID-19 severity using DERGA, a novel machine learning algorithm. *European Journal of Internal Medicine*.

Asteris, P. G., Gandomi, A. H., Armaghani, D. J., Tsoukalas, M. Z., Gavriilaki, E., Gerber, G., et al. (2024). Genetic justification of COVID-19 patient outcomes using DERGA, a novel data ensemble refinement greedy algorithm. *Journal of Cellular and Molecular Medicine*, *28*(4), Article e18105.

Asteris, P. G., Gavriilaki, E., Touloumenidou, T., Koravou, E.-E., Koutra, M., Papayanni, P. G., et al. (2022). Genetic prediction of ICU hospitalization and mortality in COVID-19 patients using artificial neural networks. *Journal of Cellular and Molecular Medicine*, *26*(5), 1445–1455.

Asteris, P. G., Kokoris, S., Gavriilaki, E., Tsoukalas, M. Z., Houpas, P., Paneta, M., et al. (2023). Early prediction of COVID-19 outcome using artificial intelligence techniques and only five laboratory indices. *Clinical Immunology*, *246*, Article 109218.

Asteris, P. G., Skentou, A. D., Bardhan, A., Samui, P., & Pilakoutas, K. (2021). Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models. *Cement and Concrete Research*, *145*, Article 106449.

Benzaamia, A., Ghrici, M., Rebouh, R., Zygouris, N., & Asteris, P. G. (2024). Predicting the shear strength of rectangular RC beams strengthened with externally-bonded FRP composites using constrained monotonic neural networks. *Engineering Structures*, *313*, Article 118192.

Chen, T.-C. T., Wu, H.-C., & Chiu, M.-C. (2024). A deep neural network with modified random forest incremental interpretation approach for diagnosing diabetes in smart healthcare. *Applied Soft Computing*, *152*, Article 111183.

Deberneh, H. M., & Kim, I. (2021). Prediction of type 2 diabetes based on machine learning algorithm. *International Journal of Environmental Research and Public Health*, *18*(6), 3317.

Farnoodian, M. E., Moridani, M. K., & Mokhber, H. (2024). Detection and prediction of diabetes using effective biomarkers. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, *12*(1), Article 2264937.

Gavriilaki, E., Asteris, P. G., Touloumenidou, T., Koravou, E.-E., Koutra, M., Papayanni, P. G., et al. (2021). Genetic justification of severe COVID-19 using a rigorous algorithm. *Clinical Immunology*, *226*, Article 108726.

Guan, H., Wang, Y., Niu, P., Zhang, Y., Zhang, Y., Miao, R., et al. (2024). The role of machine learning in advancing diabetic foot: a review. *Frontiers in Endocrinology*, *15*, Article 1325434.

Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*, 76516–76531.

Heinrichs, B., & Eickhoff, S. B. (2020). Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Human Brain Mapping*, *41*(6), 1435–1444.

Holt, R. I., & Flyvbjerg, A. (2024). *Textbook of diabetes*. John Wiley & Sons.

Hosny, M., Elshenhab, A. M., & Maged, A. (2025). Explainable AI-based method for brain abnormality diagnostics using MRI. *Biomedical Signal Processing and Control*, *100*, Article 107184.

Jain, R., Tripathi, N. K., Pant, M., Anutariya, C., & Silpasuwanchai, C. (2024). Investigating gender and age variability in diabetes prediction: A multi-model ensemble learning approach. *IEEE Access*.

Jaiswal, V., Negi, A., & Pal, T. (2021). A review on current advances in machine learning based diabetes prediction. *Primary Care Diabetes*, *15*(3), 435–443.

Kalagotla, S. K., Gangashetty, S. V., & Giridhar, K. (2021). A novel stacking technique for prediction of diabetes. *Computers in Biology and Medicine*, *135*, Article 104554.

Kannadasan, K., Edla, D. R., & Kuppili, V. (2019). Type 2 diabetes data classification using stacked autoencoders in deep neural networks. *Clinical Epidemiology and Global Health*, *7*(4), 530–535.

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *Ict Express*, *7*(4), 432–439.

Kibria, H. B., Nahiduzzaman, M., Goni, M. O. F., Ahsan, M., & Haider, J. (2022). An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. *Sensors*, *22*(19), 7268.

Kumar, S., Bhusan, B., Singh, D., & kumar Choubey, D. (2020). Classification of diabetes using deep learning. In *2020 international conference on communication and signal processing* (pp. 0651–0655). IEEE.

Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocrine Disorders*, *19*, 1–9.

Li, W., Peng, Y., & Peng, K. (2024). Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm. *PLoS One*, *19*(9), Article e0311222.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Mainali, S., Darsie, M. E., & Smetana, K. S. (2021). Machine learning in action: stroke diagnosis and outcome prediction. *Frontiers in Neurology*, *12*, Article 734345.

Mohanty, P. K., Francis, S. A. J., Barik, R. K., Roy, D. S., & Saikia, M. J. (2024). Leveraging Shapley additive explanations for feature selection in ensemble models for diabetes prediction. *Bioengineering*, *11*(12), 1215.

Nadeem, M. W., Goh, H. G., Ponnusamy, V., Andonovic, I., Khan, M. A., & Hussain, M. (2021). A fusion-based machine learning approach for the prediction of the onset of diabetes. In *Healthcare, vol. 9, no. 10* (p. 1393). MDPI.

Nimmagadda, S. M., Suryanarayana, G., Kumar, G. B., Anudeep, G., & Sai, G. V. (2024). A comprehensive survey on diabetes type-2 (T2D) forecast using machine learning. *Archives of Computational Methods in Engineering*, 1–19.

Patil, A. R., Mane, S. C., Patil, M. A., Gangurde, N. A., Rahate, P. G., & Dhanke, J. A. (2024). Artificial intelligence and machine learning techniques for diabetes healthcare: A review. *Journal of Chemical Health Risks*, 1058–1063.

Pima Indians diabetes database. (2018). URL https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data. (Accessed: 5 October 2024).

Reza, M. S., Amin, R., Yasmin, R., Kulsum, W., & Ruhi, S. (2024). Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data. *Heliyon*, *10*(2).

Rubaiat, S. Y., Rahman, M. M., & Hasan, M. K. (2018). Important feature selection & accuracy comparisons of different machine learning models for early diabetes detection. In *2018 international conference on innovation in engineering and technology* (pp. 1–6). IEEE.

Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018). Prediction of diabetes using machine learning algorithms in healthcare. In *2018 24th international conference on automation and computing* (pp. 1–6). IEEE.

Singh, N., & Singh, P. (2020). Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. *Biocybernetics and Biomedical Engineering*, *40*(1), 1–22.

Talari, P., N, B., Kaur, G., Alshahrani, H., Al Reshan, M. S., Sulaiman, A., et al. (2024). Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2. *Plos One*, *19*(1), Article e0292100.

Talebi Moghaddam, M., Jahani, Y., Arefzadeh, Z., Dehghan, A., Khaleghi, M., Sharafi, M., et al. (2024). Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm. *BMC Medical Research Methodology*, *24*(1), 220.

Tan, Y., Chen, H., Zhang, J., Tang, R., & Liu, P. (2022). Early risk prediction of diabetes based on GA-stacking. *Applied Sciences*, *12*(2), 632.

Tanim, S. A., Shrestha, T. E., Emon, M. R. I., Mridha, M., Miah, M. S. U., et al. (2025). Explainable deep learning for diabetes diagnosis with DeepNetX2. *Biomedical Signal Processing and Control*, *99*, Article 106902.

Wee, B. F., Sivakumar, S., Lim, K. H., Wong, W. K., & Juwono, F. H. (2024). Diabetes detection based on machine learning and deep learning approaches. *Multimedia Tools and Applications*, *83*(8), 24153–24185.

Yahyaoui, A., Jamil, A., Rasheed, J., & Yesiltepe, M. (2019). A decision support system for diabetes prediction using machine learning and deep learning techniques. In *2019 1st international informatics and software engineering conference* (pp. 1–4). IEEE.

Zhou, H., Xin, Y., & Li, S. (2023). A diabetes prediction model based on Boruta feature selection and ensemble learning. *BMC Bioinformatics*, *24*(1), 224.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in Genetics*, *9*, 515.