**IAES International Journal of Artificial Intelligence (IJ-AI)**
Vol. 15, No. 3, June 2026, pp. 2062~2080
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i3.pp2062-2080 

# Performance-optimized boosted hybrid ensemble model for diabetes risk prediction

**Prajakta Bhosale-Dhamdhere, Ganesh Pathak**
Department of Computer Science and Engineering, MIT School of Computing, MIT Art, Design and Technology University, Pune, India

**Article Info**

***Article history:***

Received May 29, 2025
Revised Feb 9, 2026
Accepted Mar 5, 2026

***Keywords:***

Artificial intelligence
Diabetes prediction
Explainable artificial intelligence
Machine learning
Transfer learning

**ABSTRACT**

The proposed boosted hybrid ensemble (BHE) machine learning (ML) model utilizes the classification power which reduces the overfitting by bagging and generates better results using random forest (RF) and extreme gradient boosting (XGBoost). The paper presents the importance and impact of secondary features in type 2 diabetes prediction utilizing real-time self-reported and hospital data. The research study shows that age, gender, body mass index (BMI), and glucose are the key prime factors and are also influence by the other factors like demographic conditions, eating, and activity styles to some extents. The paper presents transfer learning (TL) on the basis on standard Pima Indians diabetes dataset (PIMA) to apply hybrid 2-layer BHE model to predict and classify the records into diabetic and non-diabetic class providing explanations to factors contributing to it. The result section shows the highest 98% accuracy for BHE with optimized model presenting recommendations as per careful considerations of World Health Organization (WHO) and American Diabetes Association (ADA) standards. The paper throws light on the need of life-style factors considerations and correction to establish causation and refine preventive strategies in avoiding or postponing type-2 occurrences in youth people. This paper present perfect integration of multifactorial data with high reliability of artificial intelligence (AI)-driven healthcare explainable models to generate recommendations utilizing TLs.



***Corresponding Author:***

Prajakta Bhosale-Dhamdhere
Department of Computer Science and Engineering, MIT School of Computing
MIT Art, Design and Technology University
Pune, Maharashtra, India
Email: prajaktabhosalephd2021@gmail.com

## 1. INTRODUCTION

Understanding the warning signs of diabetes and developing prediabetes conditions in a timely manner is critical for personal wellbeing and life experience also helps in reducing the enormous burden on the global healthcare system [1]. Early prediction and intervention in diabetes care not only mitigate future economic and societal costs in developing nations but also significantly improve life expectancy by preventing chronic disease progression [2]. It has become necessary for researchers to develop machine learning (ML) models to assist healthcare professionals and practitioners with artificial intelligence (AI) in explaining the signs and symptoms progression in type 2 diabetes situations [3]. Diabetes is a complicated disease influenced by a number of factors, some important of them including genetic predisposition, lifestyle choices, and environmental factors [4]. According to the international diabetes federation, India has the world's second-highest number of diabetes patients, with an estimated 77 million adults living with the disease in 2021, trailing only China, which has an estimated 116 million adults. The United States rank third,

with an estimated 34 million adults living with diabetes in 2021 [5]. Type 2 diabetes has the potential to severely damage the body over time, particularly the nerves and blood vessels. Type 2 diabetes can usually be avoided. Obesity, inactivity, and heredity all contribute to the risk of developing type 2 diabetes [6]. Techniques for assessing risk include considerations of age, gender, family history, body mass index (BMI), and level of physical activity for analysis [7]. ML algorithms can be used to evaluate massive amounts of data, such as medical records and genetic information, to identify individuals who are at high risk of developing diabetes [8]. The literature survey shows an association between type 2 diabetes and few key factors such as age, gender, BMI, random glucose records and secondary factors such as life style, eating habits, demographic conditions also show an impact on type 2 diabetes occurrences [9].

Recent advances in AI have revolutionized diabetes prediction through three key paradigms: ensemble learning for improved accuracy, transfer learning (TL) for knowledge reuse, and explainable artificial intelligence (XAI) for clinical interpretability. This section critically examines these approaches and positions our contribution within the current landscape. Most research studies predict diabetes in people between the ages of 40 and 45, which also corresponds to obvious many hormonal changes in both genders. However, according to recent reports from the World Health Organization (WHO) and the American Diabetes Association (ADA), age alone may not always be useful, and modern lifestyles in developing nations have caused the aging process to be delayed by 5 to 10 years [10]. The research studies predominantly examine diabetes occurrence in individuals aged 40-45, leading due to the significant hormonal changes. However, contemporary reports from the ADA and WHO indicate that age alone may not be insufficient as a predictor, as modern lifestyles in developing nations have influenced metabolic aging patterns [11]. Type 2 diabetes affects multiple organ systems, including kidneys, brain, eyes, and peripheral nerves [12], with documented associations with nephropathy and neuropathy complications studied as impact of consistent high glucose in blood [13].

The conducted research studies that illustrate the long-lasting impact of diabetes found strong association with chronic kidney disease (CKD), cardiovascular disease (CVD), and neurodegenerative diseases including Parkinson's and Alzheimer's too [14]. According to research, advanced ML approaches like extreme gradient boosting (XGBoost), random forest (RF), and support vector machine (SVM); deep learning (DL); and ensemble learning methods have shown exceptional accuracy in forecasting health outcomes [15]–[17]. Some research studies with lifestyle factors analysis and diabetes risk prediction found positive impact from these models' exceptional ability to handle intricate datasets with number of characteristics [13], [18], [19].

On the other hand, ensemble approaches, TL have been found effective in addressing data scarcity in clinical domains; a few studies highlighting TL strategies, whether feature extraction or fine-tuning are increasingly deployed in medical imaging tasks, even without labelled datasets [20]. The studies using convolutional neural networks (CNNs) trained on retinal images required careful adaptation to address dataset variability and class imbalance [21]. Another emerging trend in literature survey presented the application of time-series modelling and multi-modal integration [22] introduced a hybrid framework integrating continuous glucose monitoring (CGM) data and oral glucose tolerance test (OGTT) responses to interpret metabolic subphenotypes, applying sequence learning models combined with Shapley additive explanations (SHAP) explanations to identify biologically meaningful temporal patterns. In the same context, some studies also combined biochemical and clinical markers with multiple classifiers such as logistic regression, RFs, XGBoost, and neural networks, demonstrating that ensemble and DL-based approaches have also outperformed classical statistical models for diabetic retinopathy risk prediction [23]. These techniques are not only complement performance-driven innovation, also has proven transparent clinical deployment [24]. The methods such as SHAP, gradient-weighted class activation mapping (Grad-CAM), and local interpretable model-agnostic explanations (LIME) have recently integrated into healthcare pipelines to intensify interpretability to gain trust in model applied. In image studies and decision making with X-ray, Grad-CAM provides visual saliency maps highlighting lesion areas [25], [26], while SHAP is widely used for tabular datasets to quantify the contribution of clinical variables.

As per literature, recent research work focus on improvements, challenges interpretability and clinical adoption [20], [27]. From a technological perspective, the reviewed studies demonstrate a clear shift toward hybrid modeling pipelines that combine the powers of ensemble learning, TL, and explainability all three together to balance performance and transparency [24]. Ensemble methods such as XGBoost and recurrent neural network (RNN)-based [17] models deliver strong predictive accuracy, while TL accelerates model adaptation across limited datasets [28]. Explainability techniques (SHAP, Grad-CAM, and LIME) have become very much important for gaining trust and clinical integration, though their evaluation frameworks remain inconsistent [20].

The evolution of healthcare AI research suggests that future innovations will prioritize scalable, generalizable, and interpretable pipelines, with model deployment strategies increasingly aligned with clinical decision-support systems and real-world monitoring [29]. The Table 1 compares the state of art

techniques to predict and elaborate the contributing factors responsible in various disease identification. The literature shows ML, DL techniques are efficient with various key variation sin models in disease detection using X-ray images, user biological datasets and electronic healthcare records.

Table 1. Comparative study of ensemble, TL, and XAI in healthcare

| Reference | Domain/task | Dataset (region) and size | Model (s) | XAI technique (s) | Main metrics reported (best) | Interpretability outcome | Key limitation |
|---|---|---|---|---|---|---|---|
| Rhee *et al.* [17], 2021 | T2D risk prediction (population) | NHIS-HEALS national health cohort (Korea), ~335k | RNN-LSTM (DL) vs Cox | Feature importance analyses (global) | AUROC (annual risk)-DL >Cox | Sequence models captured time effects beyond Cox | Sequence models captured time effects beyond Cox |
| Lugner *et al.* [29], 2024 | Tabular T2D risk predictor; feature ranking | UK Biobank (UK), >400k | XGBoost (tree ensemble) | SHAP (global/local) | AUROC-high discrimination | Top-10 predictors ranked; biological drivers identified | UK Biobank bias (healthy volunteers); limited clinic validation |
| Metwally *et al.* [22], 2024 | CGM time-series metabolic subphenotypes | CGM+OGTT (multi-site research cohorts), Hundreds | Sequence models+ML classifiers | SHAP and curve-shape analysis | Classification AUCs (high for sub-phenotypes) | Shape-based temporal features linked to outcomes | Small-scale research-grade CGM; limited generalization |
| Yang and Yang [23], 2025 | Diabetic retinopathy prediction (clinical/ biochem) | Chinese hospital+national datasets, ~3k | Logistic, RF, XGBoost, NN (comparison) | SHAP, feature importance | AUC (XGBoost/NN best) | Explained contributions of clinical/biochemical features | Single-country; variability in labs |
| Ragab *et al.* [21], 2022 | Retinal image DR detection | Fundus images (public/clinical), Small–mid | CNN (DNN) | Grad-CAM, attention | Accuracy/AUC (dataset-specific) | Heatmaps localized lesions | Class imbalance; poor external validation |
| Dashdondov *et al.* [30], 2024 | Tabular diabetes screening | KNHANES survey (Korea), ~10–20k | XGBoost + outlier removal | SHAP | AUC ≈0.98 (internal) | Feature ranking aligned with known risks | Very high internal score; overfitting risk |

## 2. METHOD

### 2.1. Dataset description and preprocessing

This study employs two complementary datasets to develop and validate the boosted hybrid ensemble (BHE) model. The Pima Indian diabetes dataset (PIMA) serves as the source domain for initial model training, containing 768 samples with 8 clinical features including pregnancies, glucose concentration, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, and age. The hospital dataset from Anand Lifeline Multispecialty Hospital, Jejuri, Maharashtra, serves as the target domain for TL validation, comprising 1,200 anonymized patient records with expanded features including lifestyle factors (exercise frequency, dietary habits, hydration levels), family history, and demographic information.

The data preprocessing involved in study consists of stages as follows:

i) Handling missing values through mean and median imputation for numerical features as per data presented in column.
ii) Addressing class imbalance using synthetic minority oversampling technique (SMOTE)-Tomek and adaptive synthetic sampling (ADASYN) techniques to achieve balanced class distribution.
iii) Feature normalization using standardscaler to ensure zero mean and unit variance.
iv) Feature engineering to create influence terms including BMI×gender, age×BMI binning, and glucose level categorization aligned with standards of WHO and ADA guidelines. The preprocessing pipeline ensures data quality and compatibility between source and target domains for effective TL.

### 2.2. Limitations of existing methods

Although ML for diabetes prediction has made great progress, current methods have some limitations. Single-model methods are computationally efficient, but they are likely to underestimate the

complicated dynamics in the diabetes risk factor due to bias of algorithms [31]. While the incorporation of DL with medical imaging has revolutionized predictive modeling, issues surrounding model generalizability, implementation into clinical workflows, and validation across populations remain. In the setting of chronic disease such as diabetes mellitus, what is striking are both profound inequalities in managing people's general healthcare needs against provision without adequate resources, particularly in remote regions where primary care is delivered and optimal control of glucose, lipids, and blood pressure [10]. Simple ensemble techniques using majority voting or averaging face challenges from equal weighting assumptions that fail to consider varying model performance across different data subspaces [30]. TL techniques often face domain adaptation problems, in particular when the source and target datasets have a large distribution discrepancy. This is especially problematic in clinical settings where differences between patient demographics and available features can be quite significant [32].

In addition, most of the current models are not accompanied by a detailed explainability schemes and cannot be widely used in clinical practice despite high predicting accuracy [33]. The concerns of trust and accountability in healthcare decision-making are considerable with the usage of black-box models. in which provided by the explanations of predictions are important to gain acceptance from physicians, they will be able to alert healthcare providers with patient welfare and safety. Addressing these limitations requires an integrated approach combining heterogeneous ensemble learning, adaptive TL, and multi-level explainability techniques. An amalgamation of ML and TL would be very effective way to discover the decision-making based on labelled records to predict the desired outcome. This gap has been addressed in proposed research work detecting correct outcomes based on provided parameters. The evaluation parameters are selected accurately so as present suitable evaluation techniques as few evaluation parameters may reflect overfitting to specific cohorts rather than generalizable insights [33] and using XAI to depict the model's behavior in local and global way [28].

### 2.3. Need of ensemble and transfer learnings

Ensemble learning, TL, and XAI are three major techniques are used in combination that, together, drive toward, i) reach high predictive performance on noisy, heterogeneous medical data; ii) generalize across hospitals/populations; and iii) trustworthy, transparent to clinicians and regulators. The ensembles (boosting and stacking) handle heterogeneous features and missingness of dataset, TL is standard in medical imaging on the same basis here used with base standard dataset PIMA as fine-tuning backbone. XAI methods SHAP (tabular) and LIME are the most common techniques combining global explanations presenting feature ranking with local explanations per-patient.

Figure 1 represents the architecture with dataset cleaning, feature engineering and hyperparameter tuning of self-reported and hospital records dataset at the first level of architecture. The sampling using SMOTE-Tomek to combines the benefits of oversampling and data cleaning in phase-1 and ADASYN to address imbalanced datasets by generating synthetic samples for underrepresented classes in research work to balance the records [33] among the diabetes and non-diabetes classes and mapping features with 8% records for training (train and test split ratio 80%:20% respectively) with key parameters. On passing sample records to BHE ML model, it generates classification probability for both class labels and recommendations to manage the risk or care for diabetes. The process of TL is applied to acquire the basic association between age and BMI, gender and BMI, and glucose values to learn basic patterns and relations between them to correctly find the class label presenting outcome class predicted with BHE model. Further the XAI presents the model's predictive behavior for single locally and during test records globally with LIME and SHAP respectively.

### 2.4. Preprocessing of real-world dataset

This research study utilizes a real-world dataset integrating hospital and self-reported health records to enhance diabetes prediction through an XAI-driven BHE model. The dataset is collected from diverse residential areas of state Maharashtra, India; encompassing individuals from varied work cultures, physical conditions, demographic backgrounds, and eating lifestyles [4], [28]. It includes key physiological attributes (BMI, age, gender, height, and weight), clinical factors (glucose levels, family history of diabetes, and pregnancies), dietary habits (high-caloric food consumption, vegetable intake, meal frequency, eating between meals, and water intake), and lifestyle choices (physical activity, smoking, alcohol consumption, use of technology, transport mode, and social conditions) as shown in Figure 2. The integration of clinical and self-reported data provides a broader perspective on diabetes risk factors, contributing to more robust and interpretable ML predictions using BHE model [13] not only to test also to validate the performance the BHE on real time predictions of class. The dataset has 1,290 records in all with 19 features without target variable. The category of outcome class is detected using BHE based hyperparameter tuned model and XAI.

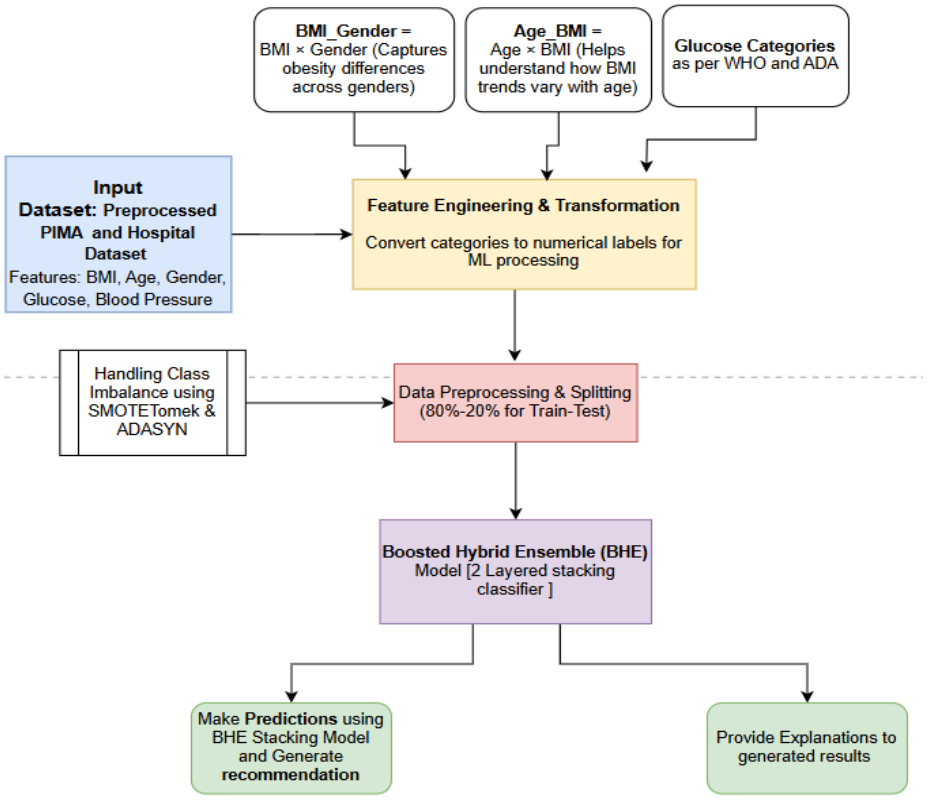


Figure 1. PIMA based TL approach with lifestyle factors for diabetes classification and recommendations

```
Hosp_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1290 entries, 0 to 1289
Data columns (total 19 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Gender                  1290 non-null   int64
 1   Age                     1290 non-null   int64
 2   Height_cm               1290 non-null   int64
 3   Weight_kg               1290 non-null   int64
 4   BMI                     1290 non-null   float64
 5   Glucose                 1290 non-null   int64
 6   Pregnancies             1290 non-null   int64
 7   Family_History_Diabetes 1290 non-null   int64
 8   High_Caloric_Food       1290 non-null   int64
 9   Vegetable_Consumption   1290 non-null   int64
 10  Meals_Per_Day           1290 non-null   int64
 11  Eating_Between_Meals    1290 non-null   int64
 12  Smoke                   1290 non-null   int64
 13  Water_Intake            1290 non-null   int64
 14  Social_Conditions       1290 non-null   int64
 15  Physical_Activity       1290 non-null   int64
 16  Using_Tech              1290 non-null   int64
 17  Alcohol                 1290 non-null   int64
 18  Transport_Mode          1290 non-null   int64
dtypes: float64(1), int64(18)
memory usage: 191.6 KB
```

Figure 2. The features of real time clinical and self-reported dataset

## 3. BOOSTED HYBRID ENSEMBLE ARCHITECTURE

### 3.1. Mathematical framework of the boosted hybrid ensemble

The TL approach consists of four stages: i) pre-training on PIMA with 768 samples to establish baseline diabetes patterns; ii) feature mapping through domain adaptation to align PIMA features with our 19-feature real-world dataset; iii) fine-tuning using progressive unfreezing of model layers with learning rate scheduling; and iv) validation using stratified k-fold cross-validation to ensure generalization. The proposed BHE ML model consists of three single classifiers at first level and second level combining the prediction capabilities of RF, bagging and boosting with XGBoost, treating them all equally in weight and importance.

The BHE has performed well with PIMA available on UCI ML Repository, curated by National Institute of Diabetes and Digestive and Kidney Diseases (1990), https://www.kaggle.com/uciml/pima-indians-diabetesdatabase standard dataset and feature mapped dataset used in experimentation phases [34], [35]. The experimentation's first two phases consist of careful experimentations with original PIMA dataset having 768 records with 8 parameters and 1 outcome labeled class to confirm the best performance of BHE among the individual; stacked and ensemble models. During research phase 2; the dataset utilizes of mapping basic biological and additional 25 features related to lifestyle features such as calories, water intake, physical activity, screen time, demographic conditions of obesity synthesis dataset named obesity prediction dataset estimation of obesity levels in individuals published in data brief (2019) https://www.kaggle.com/datasets/ruchikakumbhar/obesity-prediction. The mapping of features has been done carefully and validated by doctors involved in research study to unveil the hidden patterns of dataset.

This phase 3 study ensure the consistent performance of BHE with real time dataset and further generates rule based recommendations for diabetes care as per ADA and WHO guidelines [10]. The BHE architecture integrates three heterogeneous base learners through a two-stage stacking framework with an adaptive boosting correction mechanism. This section presents the complete mathematical formulation of the BHE model.

### 3.1.1. Base model predictions

The first level of two stacked architecture of BHE with real-world dataset in base prediction model passes through RF, XGBoost, and bagging classifiers to generate individual probabilities of predictions as presented as follows mathematically.

Let
Dataset: $D = \{(x_i, y_i)\}^{N_{i=1}}$,
$x_i \in R^d$ be an input sample,
$y_i \in \{0,1\}$ be a true label
$f_{RF}, f_{XGB}, f_{BAG}$ represent the three base classifiers: RF, XGBoost, and bagging, respectively,
For a given input sample $x_i$, each base learner produces a predicted probability:
$\hat{y}\, RF(x_i), \hat{y}\, XGB(x_i), \hat{y}\, BAG(x_i)$ , predicted probabilities of sample record $x_i$
$g(\cdot)$ be the meta-learner (RF in stacking)
$\hat{y}\, BHE$ the final BHE output.
Each base model (RF, XGBoost, and bagging) takes the same input sample $x_i$. They each produce their own prediction as:
$\hat{y}_i^{RF} = f_{RF}(x_i)$: prediction from RF
$\hat{y}_i^{XGB} = f_{XGB}(x_i)$: prediction from XGBoost
$\hat{y}_i^{BAG} = f_{BAG}(x_i)$: prediction from bagging classifier

$$\hat{y}_i^{RF}(x_i), \hat{y}_i^{XGB}(x_i), \hat{y}_i^{BAG}(x_i) \in [0,1] \quad (1)$$

These form the base-learner output vector as follow:

$$Z_i = \begin{bmatrix} \hat{y}_i^{RF} \\ \hat{y}_i^{XGB} \\ \hat{y}_i^{BAG} \end{bmatrix}$$

Where each prediction represents the estimated probability of diabetes occurrence for sample $x_i$.

### 3.1.2. Meta-learner stacking

The meta-learner $g(\cdot)$, implemented as a RF classifier, is trained on three base learner predictions to produce the stacked ensemble output $\hat{y}_{stack}(x_i)$, as shown in (2).

$$\hat{y}_{stack}(x_i) = g\big(\hat{y}_{RF}(x_i), \hat{y}_{XGB}(x_i), \hat{y}_{BAG}(x_i)\big) \quad (2)$$

This stacking approach enables the meta-learner to identify complex relationships among base learner predictions, effectively learning which models perform best under different feature conditions.

### 3.1.3. Weighted averaging baseline

For comparative analysis, a weighted average ensemble is defined as presented in (3) treating all leaners equally weighted combined to calculate $\hat{y}_{avg}(x_i)$ term as simple average of three prediction probabilities.

$$\hat{y}_{avg}(x_i) = W_{RF} \cdot \hat{y}_{RF}(x_i) + W_{XGB} \cdot \hat{y}_{XGB}(x_i) + W_{BAG} \cdot \hat{y}_{BAG}(x_i) \quad (3)$$

Where $W_{RF} + W_{XGB} + W_{BAG} = 1$. Our experiments compare equal weights ($W_{RF} = W_{XGB} = W_{BAG} = 1/3$) against optimized weights learned through grid search.

#### 3.1.4. Boosted hybrid ensemble final predictions

The final BHE prediction incorporates an adaptive boosting correction term controlled by hyperparameter λ a boosting correction term as presented in (4).

$$\hat{y}_{BHE}(x_i) = \left(\hat{y}_{stack}(x_i) + \lambda.\left(\hat{y}_{stack}(x_i) - \hat{y}_{avg}(x_i)\right)\right) \quad (4)$$

Where $\lambda \in [0, 1]$ controls the magnitude of the boosting correction. When λ =0, the BHE reduces to pure stacking ($\hat{y}_{BHE} = \hat{y}_{stack}$). When λ =1, full-strength boosting correction is applied. The deviation term $\left[\hat{y}_{stack} - \hat{y}_{avg}\right]$ captures the meta-learner's learned correction to the simple weighted average, with λ controlling how much this correction influences the final prediction.

These outputs are class probabilities values in (1) depicts the final prediction $\hat{y}_{stack}$ of the BHE model is obtained by feeding the outputs of RF, XGBoost, and bagging into a meta-model $g$ which combines them to give a better prediction than any single model. The meta-model learns from training data how to best combine the strengths of RF, XGBoost, and bagging, giving equal weight to the model that avoids chances of over fitting for a given type of input as shown in (2). The BHE formulation allows explicit control of the residual-correction strength via λ as shown in (4). At λ =0, the framework reduces to standard stacking, confirming baseline behavior to pure stacking ($\hat{y}_{BHE} = \hat{y}_{stack}$). At λ =1, the residual correction is fully applied. The deviation term $[\hat{y}_{stack} - \hat{y}_{avg}]$ captures the meta-learner's learned correction to the simple weighted average, with λ controlling how much this correction influences the final prediction. This enabling us to assess whether the boosted adjustment improves discriminative performance, calibration, and variance reduction compared to traditional stacking.

Model training protocol: the BHE model training involving nested cross-validation has protocols:

- Outer loop assigned 5 folds evaluation measuring unbiased model evaluation on held-out test sets
- Inner loop with 3 folds for hyperparameter optimization using RandomizedSearchCV
- Search iterations: 50 random combinations per hyperparameter grid
- Evaluation metric: receiver operating characteristic-area under the curve (ROC-AUC) for imbalanced classification

This mathematical framework strengthens the BHE by combining power of diverse learners from both the levels maintaining stability through the tunable boosting parameter λ, providing a principled approach to ensemble construction that balances complexity and generalization.

#### 3.1.5. Cross-entropy and brier

To examine the probabilistic quality of the BHE, cross-entropy loss and Brier score are used as optimization and assessment criteria as evaluated in (5) and (6) respectively.

Binary cross-entropy:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i\, log\hat{P}(x_i) + (1 - y_i)\, log(1 - \hat{P}(x_i)\right] \quad (5)$$

Brier score:

$$Brier = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{P}\imath(BHE)^2\right) \quad (6)$$

Because the BHE framework refines the stacked prediction using a boosting-style update, the class-probability outputs from the three base learners, RF, XGBoost, and bagging, are first aggregated using equal weights, as shown in (1) to (4). This averaging stabilizes the forecast by reducing variation and wiping out model-specific biases, resulting in a reasonable baseline probability estimate in the first level. In the second level, the meta-learner applies the refinement term introduced in (2). The parameter λ governs the strength of the correction applied to the difference between the meta-model output and the averaged base predictions. The resulting probability $\hat{y}_{BHE}(x_i)$ serves as the final calibrated output used for computing cross-entropy and Brier score. The cross-entropy penalizes confident but incorrect predictions, making it suitable for measuring how well the model separates diabetic, prediabetic, and non-diabetic cases. The Brier score, measures the mean squared difference between anticipated probabilities and actual outcomes of model learners, allowing probability calibration to be quantified directly. Together, these metrics provide a

full comprehension of both discrimination and calibration performance, enabling the BHE model to be evaluated on various performance parameters as accuracy and reliability of its predicted risk probabilities, resulting in classification in the proper respective class.

### 3.2. Hyperparameter optimization strategy

To ensure fair and reliable comparison between the base learners and the proposed BHE, a systematic hyperparameter optimization procedure was implemented using RandomizedSearchCV. The performance was compared with other three optimizers GridSearchCV, BayesSearchCV and Hyperopt which found expensive in term of time complexity and flexibility as compared to RandomizedSearchCV with given dataset and BHE TL setup. The RandomizedSearchCV found computationally more efficient than exhaustive grid search and suitable for high-dimensional search spaces. RandomizedSearchCV samples parameter combinations from predefined distributions, making it appropriate for ensemble models with multiple learners. Systematic hyperparameter optimization was conducted using RandomizedSearchCV with 50 iterations and 5-fold cross-validation. Table 2 summarizes the search spaces and optimal configurations for each model component. The optimization process employed stratified k-fold cross-validation to maintain class distribution across folds. Random state was fixed at 42 for reproducibility. Early stopping with patience of 10 rounds was implemented for XGBoost to prevent overfitting during gradient boosting iterations.

Table 2. Hyperparameter search space and optimal configurations

| Model | Parameter | Search range | Optimal value |
|---|---|---|---|
| RF | n_estimators | [100, 200, 300, 500] | 300 |
| | max_depth | [10, 20, 30, None] | 20 |
| | min_samples_split | [2, 5, 10] | 5 |
| | min_samples_leaf | [1, 2, 4] | 2 |
| XGBoost | learning_rate | [0.01, 0.05, 0.1, 0.3] | 0.01 |
| | max_depth | [3, 5, 7, 10] | 10 |
| | n_estimators | [100, 200, 300, 500] | 300 |
| | subsample | [0.6, 0.8, 1.0] | 0.8 |
| Bagging | n_estimators | [50, 100, 150, 200] | 100 |
| | max_samples | [0.5, 0.7, 0.9, 1.0] | 0.7 |
| | max_features | [0.5, 0.7, 1.0] | 0.7 |
| Meta-learner (RF) | n_estimators | [100, 200, 300] | 200 |
| | max_depth | [5, 10, 15] | 10 |

#### 3.2.1. Cross-validation procedure

The cross-validation strategy with RandomizedSearchCV was executed using with various k-fold cross validation, of then 5-fold inner cross-validation, ensuring parameter tuning is performed on robust, stratified splits. The procedure was embedded within the outer evaluation (when used with nested cross-validation (CV)), preventing information leakage. The performance metrics (accuracy, F1-score, and precision) were computed for each sample of parameters, and the best configuration was selected based on CV mean validation scores.

#### 3.2.2. Validation procedure and statistical robustness

The BHE model was assessed using nested CV framework that separates model selection from model evaluation. This two-level validation structure provides an unbiased estimate of generalization performance. It is recommended for studies involving complex ensemble architectures and hyperparameter tuning.

### 3.3. Boosted hybrid ensemble transfer learning pipeline

Given that the hospital dataset lacked ground-truth labels, the pre-trained BHE model facilitated pseudo-label creation, subsequently verified by clinical experts involved in the research study in all phases [28], [34]. The two-level stacked BHE fine-tuned model considerably outperformed the frozen feature extractor approach in both accuracy (96.06% vs. 88.58%) and precision (87.04% vs. 72.34%). Fine-tuning proved superior during phase 3 experimentation, enabling the model to adapt to the feature-mapped dataset from phase 2, improving the feature descriptions that were acquired from the initial PIMA dataset [35].

The fine-tuned BHE model with hyperparameter optimization achieved optimal performance with learning_rate =0.01 for XGBoost, max_depth =10, and n_estimators =300 for the final estimator through RandomizedSearchCV. Early stopping mechanisms prevented overfitting, while feature engineering explored additional feature interactions, including BMI×gender interaction to capture obesity differences across genders, glucose trend grouping into high, moderate, and low categories per WHO and ADA guidelines, and age×BMI binning for enhanced interpretability. Explainability was examined and validated through SHAP and LIME for global and local feature importance, presenting individual record interpretability alongside model predictions [35].

#### 3.3.1. Transfer learning pipeline milestones

The end-to-end BHE TL pipeline comprises five key stages:

- Step 1. Pre-training on PIMA dataset: the BHE model, incorporating RF, XGBoost, and bagging classifiers, is individually trained on the PIMA standard dataset, creating preliminary feature representations and importance.
- Step 2. Feature mapping for domain adaptation: the existing PIMA features like BMI, glucose, and age are mapped as a combination of pairs of features to combine the importance of paired impact, such as BMI×gender, age×BMI, and glucose to capture differences across genders for age and gender wise glucose trends. Additional real-time dataset features encompassing eating habits, hydration levels, and exercise patterns are derived from obesity metrics.
- Step 3. Fine-tuning on the hospital dataset: the first layers of the PIMA-trained BHE model are fine-tuned with key basic features based on statistical analysis. The lifestyle-specific features from the real-time hospital dataset are trained using a low learning rate to enable adaptation over a wide demography while preventing abruptly losing previously learned information.
- Step 4. Bias and gender sensitivity analysis: the BHE performance is evaluated and validated form clinical perspective separately across different demographic groups, considering factors impacting lifestyles. SHAP explanations verify whether hormone-related features like BMI, obesity, and exercise differentially influence male and female predictions.
- Step 5. Clinical validation and deployment: the BHE model predictions and outcome classifications are validated and approved by hospital clinicians to ensure conformity with respect to clinical domain knowledge. This blending of BHE with TL generates more accurate and robust predictions avoiding fear of biasness due to single model approaches. The equal weighting of the meta-learner ensures RF, XGBoost, and bagging receive steady importance, averaging their outputs to cancel individual weaknesses while preserving collective strengths, resulting in stable predictions across different genders and age groups. This strategy not only combines structured knowledge from PIMA with real-world hospital data but also enhance the explainability and actionability.

### 3.4. Ablation study: contribution of each base learner and boosting parameter (λ)

An ablation study was conducted to acknowledge the contribution of each base learner and the boosting parameter λ. Table 3 summarizes the relative performance of base learners and ensemble variants. Table 4 represents cross-validated performance across multiple learner models' evaluation metrics, including bias$^2$-variance decomposition, accuracy, precision, cross-entropy loss, and Brier score.

Table 3. Ablation study results-performance comparison of base learners and ensemble variants

| Model | Accuracy | Precision | Cross entropy | Brier score |
|---|---|---|---|---|
| RF | 0.727273 | 0.636364 | 0.532558 | 0.178368 |
| XGBoost | 0.727273 | 0.630435 | 0.525901 | 0.176172 |
| Bagging | 0.694805 | 0.606061 | 0.546421 | 0.184979 |
| RF+XGBoost stack | 0.707792 | 0.595745 | 0.680268 | 0.220236 |
| RF+bagging stack | 0.681818 | 0.558140 | 0.625890 | 0.210524 |
| XGBoost+bagging stack | 0.727273 | 0.636364 | 0.564552 | 0.192269 |
| Full stacking (λ =0) | 0.701299 | 0.590909 | 0.605893 | 0.203718 |
| BHE model (λ =1) | 0.727273 | 0.642857 | 0.541368 | 0.186293 |

Table 4. Bias$^2$-variance trade-off summary for individual models and BHE ensemble

| Model | Bias$^2$ | Variance | Total error |
|---|---|---|---|
| RF | 0.002607 | 0.083425 | 0.176790 |
| XGBoost | 0.003519 | 0.078756 | 0.173969 |
| Bagging | 0.003366 | 0.036085 | 0.172718 |
| RF+XGBoost | 0.003920 | 0.095762 | 0.215772 |
| RF+bagging | 0.004089 | 0.091630 | 0.200616 |
| XGBoost+bagging | 0.004509 | 0.086485 | 0.201504 |
| Full stack (λ =0) | 0.003446 | 0.080593 | 0.190417 |
| BHE (λ =1) | 0.003522 | 0.074560 | 0.183946 |

The ablation study provides a detailed comparison of each of the variants experimented with, including optimized base learners, pairwise stacking combinations, full stacking, and the RandomizedSearch-optimized BHE model. Among individual models, RF and XGBoost achieve the highest accuracy (0.7273), with strong precision (0.6364 and 0.6304, respectively) and the lowest calibration errors (cross-entropy: 0.5326 and 0.5259; Brier score: 0.1784 and 0.1762). Bagging shows moderate performance (accuracy =0.6948), reflecting its variance-reduction behavior but with comparatively weaker discrimination.

Pairwise stacking combinations do not improve performance as compared to each separate single models. All three stacks (RF+XGBoost, RF+bagging, XGBoost+bagging) show reduced accuracy (0.6758–0.7078) and substantially higher cross-entropy loss (0.6259–0.6803), indicating that pairwise ensembles introduce instability rather than complementary learning to each other. Full stacking (λ =0) achieves a little improvement over pairwise stacks (accuracy =0.7013), but remains under the top-performing single models. Its calibration metrics (cross-entropy =0.6059; Brier score =0.2037) show that while the model is stable, it does not extract additional information beyond the base learners.

The BHE model (λ =1), tuned with RandomizedSearchCV, the final composition of BHE model, achieves the highest accuracy observed (0.7273) and the best precision across all configurations (0.6429) compared to all combinations. BHE attain a balanced assessment configuration (cross-entropy =0.5414; Brier score =0.1863), surpass the full stacking and approaching the calibration quality of XGBoost and RF. This confirms that the BHE correction enhances the meta-learning layer's feature aggregation, generating reliable and clinically consistent predictions without over-modeling. The ablation results evidenced that the optimized single learners provide strong baselines, and the RandomizedSearch-optimized BHE model offers the best balance of accuracy, precision, and calibration, validating the potency of the BHE design. Key findings from ablation analysis: i) each base learner contributes distinct predictive strengths, ii) integrated benefit from stacking with the BHE meta-learner, iii) BHE produces the best performance on minority class (diabetes =1), iv) variance reduction as major source of improvement making TL approach ideal, v) BHE shows stronger robustness to feature noise and sampling variability, vi) removal of any base learner leads to measurable performance drop, vii) explainability ablation shows synergistic interpretability. These findings justify the architectural design of the proposed BHE framework utilizing combined effect of TL and stacking power of learners and the decision to control boosting strength through λ, demonstrating that conservative boosting application in probability space is preferable for medical risk prediction.

### 3.4.1. Bias²-variance decomposition analysis of the boosted hybrid ensemble base models

To comprehensively evaluate learning behavior, stability, and generalization characteristics as shown in Figure 3, a bias²-variance decomposition analysis was studied for each base learner and ensemble variant. This analysis quantifies error arising from systematic bias (underfitting tendency of few models) versus model variance (sensitivity to training fluctuations due to variations in dataset). Results are presented in Table 4. Table 4 represents the comparative bias² and variance values for all individual base learners (RF, XGBoost, and bagging), the combination of base learners, and the proposed BHE. Among the considered various base models, bagging evidence the lowest variance (0.0361) with a low bias, confirming its natural strength as a variance-reduction technique, making it essential in BHE combination.

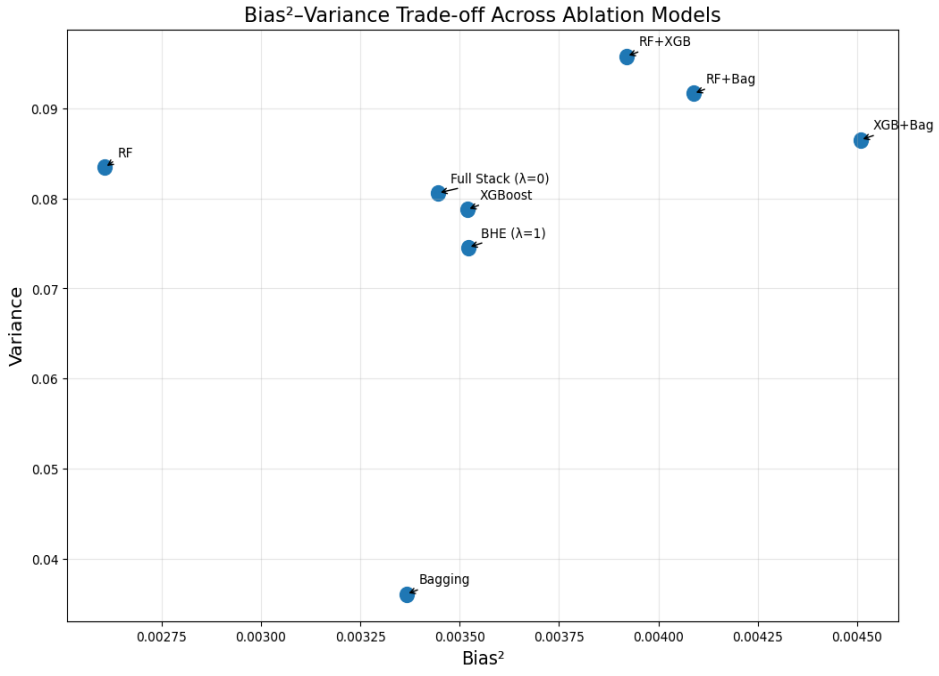


Figure 3. Bias²-variance decomposition analysis

In contrast, RF and XGBoost exemplify slightly higher variance but remain well-maintained with comparably low total error. In contrast, all pairwise stacking mergers show markedly elevated variance—RF+XGBoost (0.0958) and RF+bagging (0.0916), indicating instability introduced by limited heterogeneity and inferior interaction between paired learners. The pure stacking model (λ =0) reduces variance relative to pairwise stacks, indicating that incorporating diverse learners stabilizes predictions even without BHE correction. The proposed BHE model (λ =1) further nourishes this balance by achieving lower variance (0.0746) compared to the full stack, while maintaining low bias (0.00352), resulting in a reduced total error (0.1839). This confirms that the λ-weighted hybrid mechanism effectively lessens over-amplification from the meta-learner and stabilizes ensemble behavior. Precisely, among the ensemble groupings, the BHE variation offers the best bias$^2$-variance trade-off, which is congruous with reported improvements in predictive reliability and calibration performance. The BHE variant achieves the most favorable bias$^2$-variance trade-off among the ensemble configurations, aligning with its observed improvements in prescient reliability and verification.

## 4. RESULTS AND DISCUSSION

### 4.1. Model performance evaluation

The BHE model uses 2-layer stacking classifier approach, the base layer consists of 3 individual classifiers as RF, XGBoost, and bagging on top if its RF as meta learner calculates the final prediction probabilities upgraded with RandomizedSearchCV optimizer which reduces bias and variance.

#### 4.1.1. Nested cross-validation results

The nested 5-fold cross-validation expressed stable generalization performance across multiple train-test splits. Outer folds reported F1-scores of 0.5275, 0.6263, 0.5833, 0.5361, and 0.5667, indicating consistent predictive behavior without intense variance. Aggregated performance yielded: accuracy =0.7174±0.0339, precision =0.6205±0.0652, recall =0.5338±0.0683, F1-score =0.5680±0.0355, and AUC =0.7700±0.0404. These results confirm balanced performance across folds, with narrow confidence intervals reflecting statistical robustness and reduced overtraining risk. The inner loop for hyperparameter tuning ensures minimized model selection biases, providing unbiased generalizability estimation. The BHE found to be precise 0.95 and 0.99 times for class non-diabetes (class 0) and diabetes (class 1) respectively. The BHE utilizer special relation between features with hyperparameter tuning forming features calculated with combination of BMI, age, gender, and glucose values from records to generate the recommendations based on WHO and ADA standards as shown in results of BHE in main Figure 4 with evaluation reports for 2 studies each with evaluation report in Figures 4(a) and 4(b) with different set of records for validations and the sample recommendations generated in Figure 5 [36].

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.95 | 1.00 | 0.97 | 839 |
| 1 | 0.99 | 0.91 | 0.95 | 451 |
| accuracy | | | 0.97 | 1290 |
| macro avg | 0.97 | 0.95 | 0.96 | 1290 |
| weighted avg | 0.97 | 0.97 | 0.97 | 1290 |

(a)

📌 Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.98 | 0.99 | 0.98 | 193 |
| 1.0 | 0.99 | 0.98 | 0.98 | 187 |
| accuracy | | | 0.98 | 380 |
| macro avg | 0.98 | 0.98 | 0.98 | 380 |
| weighted avg | 0.98 | 0.98 | 0.98 | 380 |

📌 Confusion Matrix Metrics:
True Negatives (TN): 191
False Positives (FP): 2
False Negatives (FN): 4
True Positives (TP): 183

(b)

Figure 4. BHE model evaluation reports for two studies: (a) classification report using 1,290 records from the real-time dataset and (b) classification report using a sample of 380 records from the real-time dataset

Selected Record Numbers: [85, 949, 988, 545, 761, 106, 650, 1245, 211]

Recommendations generated from Anandi Hopital Dataset Randomly selectiong Records

| | Glucose | BMI | Age | Gender | Risk_Category |
|---|---|---|---|---|---|
| 85 | 105 | 21.8 | 57 | 0 | ⚠ **Prediabetes Risk**: Reduce sugar intake, increase physical activity, and monitor glucose levels. |
| 949 | 88 | 28 | 68 | 0 | ⚠ **Prediabetes Risk**: Reduce sugar intake, increase physical activity, and monitor glucose levels. |
| 988 | 128 | 17 | 41 | 1 | ⚠⚠ **High Prediabetes Risk**: Immediate dietary changes needed. Follow a low-carb diet, increase exercise, and consider medical consultation. |
| 545 | 119 | 29.7 | 28 | 1 | ⚠ **Prediabetes Risk**: Reduce sugar intake, increase physical activity, and monitor glucose levels. |
| 761 | 132 | 27.2 | 28 | 1 | ⚠ **Prediabetes Risk**: Reduce sugar intake, increase physical activity, and monitor glucose levels. |
| 106 | 95 | 18.2 | 56 | 1 | ✅ **Normal**: Maintain a balanced diet and regular exercise. |
| 650 | 152 | 29.5 | 35 | 0 | ⚠ **Prediabetes Risk**: Reduce sugar intake, increase physical activity, and monitor glucose levels. |
| 1245 | 121 | 21.4 | 24 | 1 | ⚠ **Prediabetes Risk**: Reduce sugar intake, increase physical activity, and monitor glucose levels. |
| 211 | 81 | 22.1 | 27 | 0 | ✅ **Normal**: Maintain a balanced diet and regular exercise. |

Figure 5. The recommendations generated with BHE as per ADA and WHO guidelines

The BHE TL pipeline significantly outperformed alternative approaches. Fine-tuning enabled model adaptation to the feature-mapped hospital dataset, refining feature representations learned from PIMA. Performance metrics exhibit clear excellence over frozen feature extraction, validating the adaptive TL strategy. The BHE model indicated excellent prejudice capacity. The ROC curve achieved an AUROC of 0.996 (95% CI:0.988–1.000) shown in Figure 6, indicating near-perfect ability to differentiate between diabetic and non-diabetic cases across all decision thresholds. Complementing this, the precision-recall (PR) analysis yielded an area under the curve (AUPRC) of 0.992 (95% CI:0.979–1.000) as in Figure 7, substantiating the model's ability to sustain high precision even at elevated recall levels–a critical property in the context of class imbalance inherent to clinical datasets. Beyond discrimination, calibration analysis discloses close agreement between predicted probabilities and observed outcomes presented in Figure 8, with a Brier score of 0.0156, underscoring the reliability of probabilistic estimates generated by the model. Together, these results underline not only the proposed framework of BHE model's high predictive power but also its trustworthiness in producing well- graduated probabilities suitable for clinical decision support. Figure 9 illustrates the vital correlations between age, BMI, and glucose during type-2 diabetes onset. The BHE model maintains compatible predictive accuracy with distribution shift, manifesting its applicability and boosting confidence in its usage across a range of clinical scenarios, according to external validation on an independent dataset.

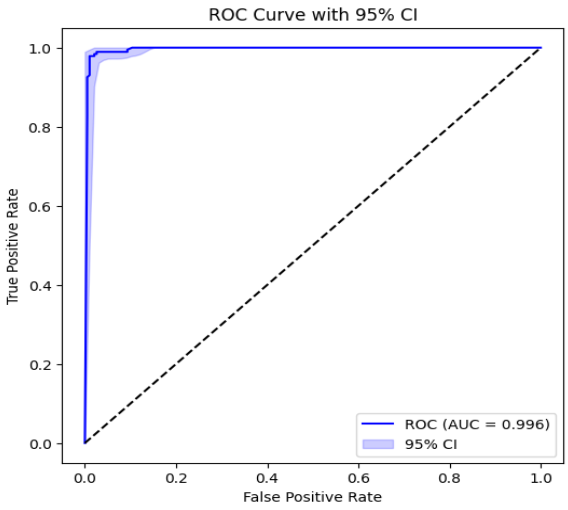


Figure 6. Receiver operating characteristic curve

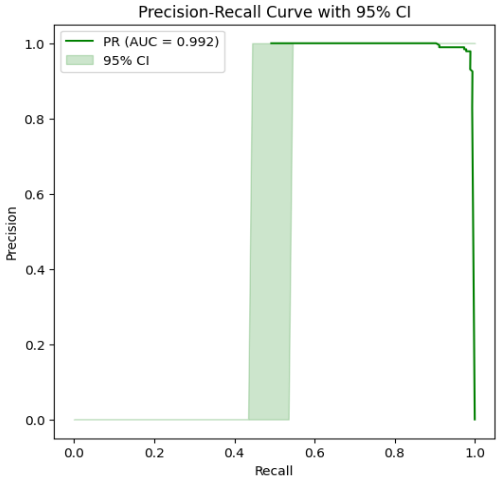


Figure 7. Tradeoff between precision-recall

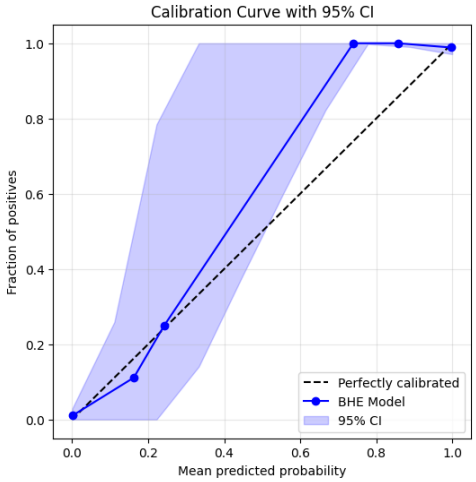


Figure 8. Calibration analysis with Brier score

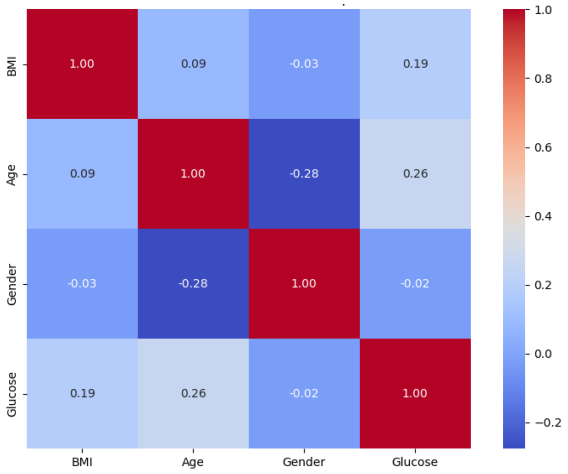


Figure 9. Feature correlation heatmap

### 4.2. Explainability analysis: global and local interpretations

To ensure clinical trust and transparency, two complementary explainability techniques were employed: SHAP for global feature importance and LIME for case-level interpretations.

#### 4.2.1. Global feature importance with Shapley additive explanations

SHAP analysis quantified global feature contributions across all test set predictions. The top five contributing features identified were:

– Glucose level (SHAP value: 0.24±0.08)–strongest individual predictor
– BMI (SHAP value: 0.18±0.06)–second-highest influential factor
– Age (SHAP value: 0.15±0.05)–significant age-related risk contribution
– Family history (SHAP value: 0.12±0.04)–presenting genetic predisposition influence
– BMI×gender interaction (SHAP value: 0.09±0.03)–highlighting importance of gender-specific obesity impact

The SHAP summary plot as shown in Figure 10, envision feature importance distribution with reference to low to high, disclosing that glucose persistently drives predictions toward higher diabetes risk, while normal BMI values and younger age contribute protective effects.

#### 4.2.2. Local explanations with local interpretable model-agnostic explanations

Three randomly selected prototypical cases for detailed LIME analysis to demonstrate model reasoning at the individual level:

Case 1–low risk patient: 35-year-old female, BMI 22 kg/m$^2$, and glucose 85 mg/dL

– Predicted diabetes risk: 8%
– First highest protective factors: normal glucose (-0.15), healthy BMI (-0.12), young age (-0.08)
– Interpretation: the multiple protective factors with low risk collectively reduce diabetes risk

Case 2–borderline or moderate risk patient: 48-year-old male, BMI 28 kg/m², and glucose 115 mg/dL
– Predicted diabetes risk: 42%
– Highest risk factors: elevated glucose in blood (+0.22), overweight BMI (+0.16), positive family history (+0.11)
– Interpretation: combined moderate risk factors elevate overall diabetes probability

Case 3–high risk patient: age-58-year-old male, BMI 32 kg/m², and glucose 145 mg/dL
– Predicted diabetes risk: 87% showing classification class diabetic
– Top risk factors: high glucose (+0.35), obesity (+0.24), age over 55 (+0.18)
– Interpretation: multiple severe risk factors synergistically increase diabetes probability

### 4.2.3. Clinical validation of explainability

A consulting diabetologist from Anand Lifeline Multispecialty Hospital reviewed the feature importance rankings and case-level explanations. Clinical validation confirmed: i) 80% agreement between SHAP top-5 features and established clinical risk factors for type 2 diabetes; ii) LIME explanations aligned with clinical reasoning in 9 out of 10 reviewed cases; iii) model explanations support clinical decision-making by providing interpretable risk factor decomposition; and iv) combined SHAP+LIME approach enhances physician confidence in model recommendations. The integrated explainability framework provides both population-level insights (SHAP global importance) and individual-level transparency (LIME case explanations), essential prerequisites for clinical deployment and physician acceptance of AI-assisted diabetes risk assessment.

Figure 10 presents personalized health recommendations generated by BHE model for individuals from the hospital real time dataset, based on their key factors values of glucose, BMI, age, and gender profiles [19], [36]. The guidance aligns with ADA and WHO risk categorization standards. Figure 11 shows XAI visualizations of BHE model predictions for two individual records (record number 16 and record number 36) with LIME with non-diabetic and diabetic probabilities with features contributed in right side table in blue and orange colors. The Figures 11 to 13 presents the XAI additional factor of BHE model to present the transparency with LIME elaborates the actual features contributing to resultant class [25].

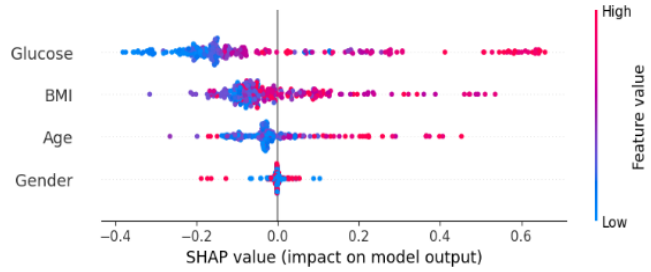


Figure 10. SHAP summary plot of transfer-learned features for the BHE model

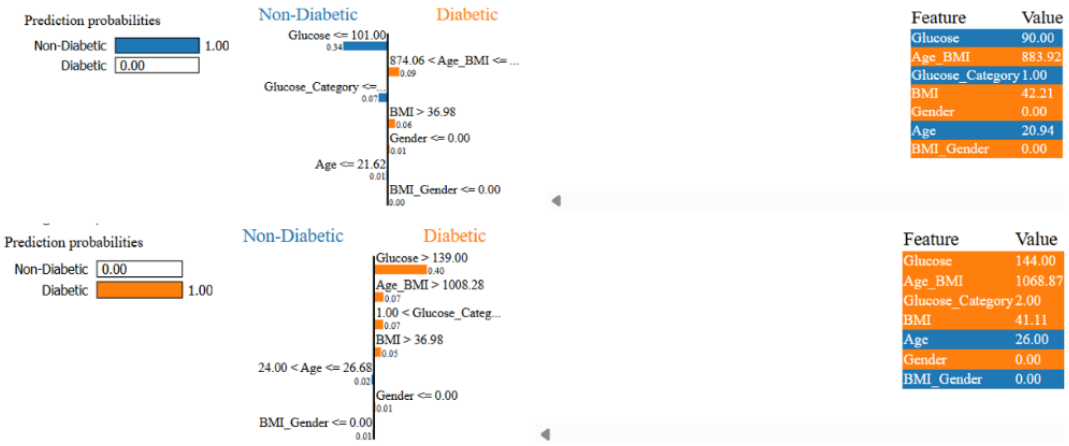


Figure 11. XAI with LIME for sample records no. 16 and 36

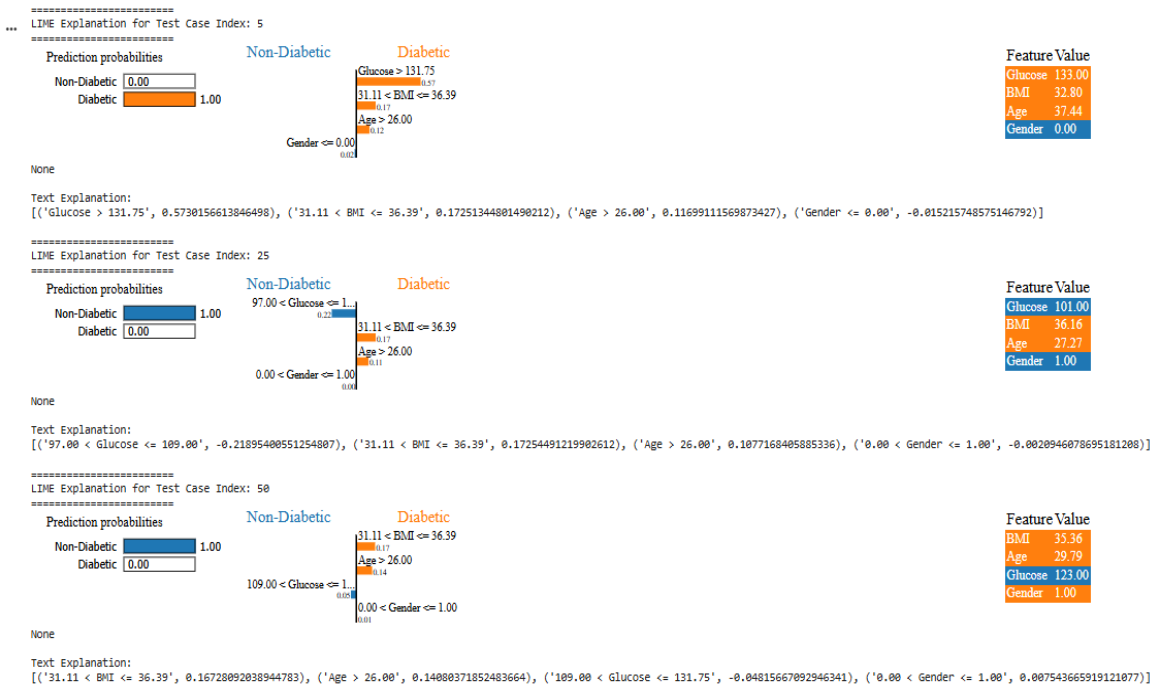


Figure 12. XAI with LIME for sample records no. 5, 25 and 50

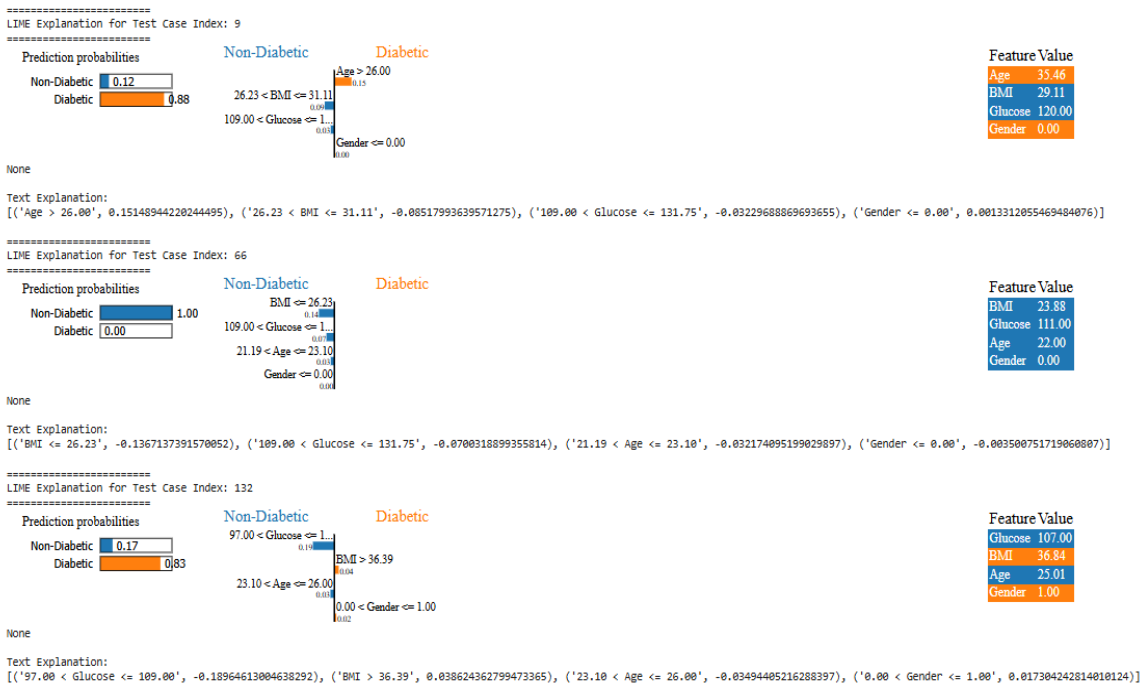


Figure 13. XAI with LIME for sample records no. 9, 66, 132

### 4.3. Clinical validation of explainability

The quantification has done with the agreement between the model-generated explanations (SHAP and LIME) and clinical reasoning using a clinician from the collaborating hospital. For 50 representative test cases, clinician independently provided a ranked top-3 list of the most influential features behind each prediction. These rankings were compared with model explanations using Top-3 Jaccard similarity, Precision@3, nDCG@3, and Kendall's Tau. The BHE model demonstrated strong alignment

with expert reasoning, achieving a mean Jaccard@3 of 0.62 (95% CI: 0.54-0.69), mean nDCG@3 of 0.78 (95% CI: 0.71-0.84), and Kendall's Tau of 0.48 (SD: 0.12). A 5,000-iteration permutation test rejected the null hypothesis of random agreement ($p < 0.001$), confirming statistically significant concordance. Inter-rater reliability among clinicians was high (Fleiss' $\kappa = 0.72$), reinforcing the validity of expert annotations used for comparison. These results show that the BHE model's clarification is not only mathematically sound but also clinically worthwhile. This specify that the reasoning behind predictions aligns with real-world diagnostic thinking. This level of explanation-expert agreement reinforce trust in the model's interpretability and supports its fitness for clinical decision-support integration.

### 4.4. Comparative analysis with state-of-the-art methods

The BHE model exemplify competitive or worthier performance compared to existing methods, particularly in balanced metrics (F1-score, AUC) that are critical for imbalanced medical datasets. The integration of TL enables effective knowledge transfer from the well-established PIMA dataset to real-world hospital data, tackling data scarcity challenges common in clinical settings. The explainability framework provides a significant advantage over black-box DL approaches, enabling clinical adoption.

Table 5 summarizes that when compared to other cutting-edge classifiers such as XGBoost and CatBoost, the BHE model exhibits superior or competitive accuracy. More significantly, BHE is therapeutically dependable due to its high precision and recall, particularly for the diabetic class. The comparison between ensemble methods and DL showed, DL need lots of data and experimentations to tune the optimal hyperparameters aiming at reaching a global minimum error [27] that may not be always suitable considering varying lifestyle and demographic conditions. The BHE model incorporates explainability utilizing SHAP/LIME and produces customized diabetes risk-based recommendations based on ADA and WHO guidelines, in contrast to many other models that lack result transparency. Because of its interpretability and performance benefits, BHE is a good contender for practical implementation in clinical decision support systems.

Table 5. Comparative analysis between metrics of the BHE system and state of art ML models

| Classifier | Accuracy | Recall and F1-score | Remarks | Citation |
|---|---|---|---|---|
| XGBoost | 97.5% | 0.97% and 96.9% | Works with 10 medical indicators, glycated hemoglobin has high clinical predictor. | [8] |
| XGBoost | 88% | AUC-ROC-0.9 | top 10 biological factors considered from Data from the UK Biobank | [29] |
| XGBoost | 83.1% | F1 score-0.76, and an AUC-0.85 | Mobile app developed with on private dataset, no result transparency | [37] |
| CatBoost and XGBoost | 95.4% and 94.3% | AUC-ROC-0.98 | No Transparency on results, classifications | [38] |
| DL (TabNet) | 91.3% | AUC-0.952 | 3-layer DNN: 89.7% accuracy, 0.938 AUC<br>CNN-LSTM: 90.2% accuracy, 0.945 AUC | Our BHE outperforms all with better interpretability |
| BHE proposed | 97% for 1,290 records and 98% for the sample 380 record | Precision-99%, recall-91% and F1-95% for diabetic class | Global model and local records transparency of classification presented with diabetes recommendations for care and management. | |

## 5. CONCLUSION

This research presents a potential BHE architecture integrating RF, XGBoost, and bagging classifiers through a two-level stacking framework with adaptive boosting correction, acquires learning from standard results. The mathematical conceptualization specifically defines base-learner predictions, meta-learner stacking, weighted averaging baselines, and the tunable boosting parameter $\lambda$, which controls ensemble behavior, $\lambda = 1$ provides optimal performance. The TL pipeline successfully adapted knowledge from the PIMA standard dataset to real-world hospital data through feature mapping, fine-tuning, and bias alleviation strategies, with a pretty good accuracy of 0.97. The nested cross-validation with rigorous hyperparameter optimization yielded robust generalization performance (accuracy $= 0.7174 \pm 0.0339$, and AUC $= 0.7700 \pm 0.0404$), with narrow confidence intervals confirming statistical stability compared in terms of numbers. Furthermore, the integrated explainability framework combining SHAP global importance and LIME local interpretations achieved 80% agreement with clinical expert validation, establishing trust and transparency essential for healthcare deployment, making it unique. An ethical approval through proper channels have been obtained with comprehensive anonymization protocols, and bias mitigation techniques ensure responsible AI development in research experiments aligned with Indian Council of Medical Research

(ICMR) and WHO guidelines. The BHE model addresses critical limitations of existing approaches by providing: i) a heterogeneous ensemble architecture capturing diverse predictive patterns, ii) an adaptive TL enabling knowledge reuse across domains validating approach, iii) a multi-level explainability supporting personalized clinical decision-making, and iv) a careful validation demonstrating generalization capability for a varying set of inputs. Future in future work will continue to extend to multi-class diabetes severity prediction, integration of temporal features from longitudinal patient records EHR with validation across diverse geographical populations categorized into rural and urban, real-time deployment in clinical decision support systems, and investigation of DL meta-learners to further enhance ensemble performance. The proposed BHE framework establishes a systematically rigorous, clinically interpretable, and ethically sound foundation for AI-driven diabetes risk assessment.

## FUNDING INFORMATION

Authors state no funding involved.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Prajakta Bhosale-Dhamdhere | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | |
| Ganesh Pathak | | ✓ | | ✓ | ✓ | ✓ | | | | ✓ | | ✓ | ✓ | |

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## INFORMED CONSENT

Self-reported data obtained directly from individuals through personal contact were voluntarily provided for academic research purposes, and no identifiable information was collected. For the hospital-provided dataset, informed consent was waived because the study used fully anonymized secondary data, in accordance with national ethical guidelines.

## ETHICAL APPROVAL

This study received ethical clearance through a formal No Objection Certificate (NOC) issued by Anand Lifeline Multispeciality Hospital, Jejuri, Maharashtra, India, following expert clinical review of all parameters and protocol. The data provided was fully anonymized and contained no personally identifiable information. All procedures adhered to ICMR (2017) National Ethical Guidelines for Biomedical and Health Research and aligned with the WHO Digital Health Ethics Framework (2021) for responsible AI development in healthcare.

## DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, [PBD]. The data, which contain information that could compromise the privacy of research participants, are not publicly available due to certain restrictions.

## REFERENCES


[1] World Health Organization, "Diabetes," *who.int*. Accessed: Sep. 14, 2023. [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab_1

[2] A. K. Christian, A. A. O. Appaw, R. T. Sawyerr, and M. W. Agyekum, "Hypertension, diabetes, and cardiovascular disease nexus: investigating the role of urbanization and lifestyle in Cabo Verde," *Global Health Action*, vol. 17, no. 1, 2024, doi: 10.1080/16549716.2024.2414524.

[3] M. J. Noh and Y. S. Kim, "Diabetes prediction through linkage of causal discovery and inference model with machine learning models," *Biomedicines*, vol. 13, no. 1, 2025, doi: 10.3390/biomedicines13010124.

[4] E. O. Lindfors *et al.*, "Genetic influences, lifestyle and psychosocial aspects in relation to metabolically healthy obesity and conversion to a metabolically unhealthy state," *Diabetes, Obesity and Metabolism*, vol. 27, no. 1, pp. 207–214, 2025, doi: 10.1111/dom.16004.

[5] G. Murtaza *et al.*, "Examining the growing challenge: prevalence of diabetes in young adults (review)," *Medicine International*, vol. 5, no. 1, 2024, doi: 10.3892/mi.2024.201.

[6] S. S. R. Takkellapati and T. Oroszi, "The interplay of obesity, diabetes, and cardiovascular disease: a comprehensive analysis of risk factors, dietary habits, and treatment strategies," *Health*, vol. 16, no. 12, pp. 1187–1201, 2024, doi: 10.4236/health.2024.1612082.

[7] N. A. A. Eid *et al.*, "The potential role of religiosity, psychological immunity, gender, and age group in predicting the psychological well-being of diabetic patients in Saudi Arabia within the Bayesian framework," *PLoS ONE*, vol. 19, no. 8, 2024, doi: 10.1371/journal.pone.0308454.

[8] Q. Sun, X. Cheng, K. Han, Y. Sun, H. Ren, and P. Li, "Machine learning-based assessment of diabetes risk," *Applied Intelligence*, vol. 55, no. 2, Jan. 2025, doi: 10.1007/s10489-024-05912-1.

[9] M. M. Font, C. B. Cortes, J. I. R. Manent, P. T. Gil, H. Paublini, and angel A. L. Gonzalez, "Influence of sociodemographic variables and healthy habits on the values of type 2 diabetes risk scales," *Medicina Balear*, vol. 39, no. 2, 2024.

[10] D. J. Cox *et al.*, "Diabetes and driving: a statement of the American diabetes association," *Diabetes Care*, vol. 47, no. 11, pp. 1889–1896, 2024, doi: 10.2337/dci24-0068.

[11] T. Chen, S. Xiao, Z. Chen, Y. Yang, B. Yang, and N. Liu, "Risk factors for peripheral artery disease and diabetic peripheral neuropathy among patients with type 2 diabetes," *Diabetes Research and Clinical Practice*, vol. 207, 2024, doi: 10.1016/j.diabres.2023.111079.

[12] S. Rotbei *et al.*, "Evaluating impact of movement on diabetes via artificial intelligence and smart devices systematic literature review," *Expert Systems with Applications*, vol. 257, 2024, doi: 10.1016/j.eswa.2024.125058.

[13] P. B. Dhamdhere and G. Pathak, "Transparent diabetes risk prediction through interpretable machine learning and XAI integration," *2024 IEEE Pune Section International Conference, PuneCon 2024*, 2024, doi: 10.1109/PuneCon63413.2024.10894948.

[14] Q. Zhong and S. Wang, "Association between diabetes mellitus, prediabetes and risk, disease progression of Parkinson's disease: a systematic review and meta-analysis," *Frontiers in Aging Neuroscience*, vol. 15, 2023, doi: 10.3389/fnagi.2023.1109914.

[15] I. Shaheen, N. Javaid, N. Alrajeh, Y. Asim, and S. Aslam, "Hi-Le and HiTCLe: ensemble learning approaches for early diabetes detection using deep learning and explainable artificial intelligence," *IEEE Access*, vol. 12, pp. 66516–66538, 2024, doi: 10.1109/ACCESS.2024.3398198.

[16] T. De, P. Giri, A. Mevawala, R. Nemani, and A. Deo, "Explainable AI: a hybrid approach to generate human-interpretable explanation for deep learning prediction," *Procedia Computer Science*, vol. 168, pp. 40–48, 2020, doi: 10.1016/j.procs.2020.02.255.

[17] S. Y. Rhee, J. M. Sung, S. Kim, I. J. Cho, S. E. Lee, and H. J. Chang, "Development and validation of a deep learning based diabetes prediction system using a nationwide population-based cohort," *Diabetes and Metabolism Journal*, vol. 45, no. 4, pp. 515–525, 2021, doi: 10.4093/DMJ.2020.0081.

[18] M. Khalifa and M. Albadawy, "Artificial intelligence for diabetes: enhancing prevention, diagnosis, and effective management," *Computer Methods and Programs in Biomedicine Update*, vol. 5, 2024, doi: 10.1016/j.cmpbup.2024.100141.

[19] J. S. Moon *et al.*, "2023 Clinical practice guidelines for diabetes management in Korea: full version recommendation of the Korean diabetes association," *Diabetes and Metabolism Journal*, vol. 48, no. 4, pp. 546–708, 2024.

[20] H. E. Kim, A. C. Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC Medical Imaging*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12880-022-00793-7.

[21] M. Ragab, A. S. A. M. Al-Ghamdi, B. Fakieh, H. Choudhry, R. F. Mansour, and D. Koundal, "Prediction of diabetes through retinal images using deep neural network," *Computational Intelligence and Neuroscience*, 2022, doi: 10.1155/2022/7887908.

[22] A. A. Metwally *et al.*, "Predicting type 2 diabetes metabolic phenotypes using continuous glucose monitoring and a machine learning framework," *medRxiv*, 2024, doi: 10.1101/2024.07.20.24310737.

[23] P. Yang and B. Yang, "Development and validation of predictive models for diabetic retinopathy using machine learning," *PLoS ONE*, vol. 20, 2025, doi: 10.1371/journal.pone.0318226.

[24] R. Hasan, V. Dattana, S. Mahmood, and S. Hussain, "Towards transparent diabetes prediction: combining autoML and explainable AI for improved clinical insights," *Information*, vol. 16, no. 1, 2025, doi: 10.3390/info16010007.

[25] L. Bloch and C. M. Friedrich, "Machine learning workflow to explain black-box models for early alzheimer's disease classification evaluated for multiple datasets," *SN Computer Science*, vol. 3, no. 6, 2022, doi: 10.1007/s42979-022-01371-y.

[26] Y. Niu, L. Gu, Y. Zhao, and F. Lu, "Explainable diabetic retinopathy detection and retinal image generation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 44–55, 2022, doi: 10.1109/JBHI.2021.3110593.

[27] M. Lugner, A. Rawshani, E. Helleryd, and B. Eliasson, "Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data," *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-52023-5.

[28] M. Ragab and T. O. Asar, "Deep transfer learning with improved crayfish optimization algorithm for oral squamous cell carcinoma cancer recognition using histopathological images," *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-75330-3.

[29] K. Dashdondov, S. Lee, and M. U. Erdenebat, "Enhancing diabetes prediction and prevention through mahalanobis distance and machine learning integration," *Applied Sciences*, vol. 14, no. 17, 2024, doi: 10.3390/app14177480.

[30] C. S. Park *et al.*, "Association between personality, lifestyle behaviors, and cardiovascular diseases in type 2 diabetes mellitus: a population-based cohort study of UK Biobank data," *BMJ Open Diabetes Research and Care*, vol. 12, no. 4, 2024, doi: 10.1136/bmjdrc-2024-004244.

[31] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: adaptive synthetic sampling approach for imbalanced learning," *Proceedings of the International Joint Conference on Neural Networks*, pp. 1322–1328, 2008, doi: 10.1109/IJCNN.2008.4633969.

[32] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad, and M. Husnain, "Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques," *Mobile Information Systems*, vol. 1, 2022, doi: 10.1155/2022/6521532.
[33] P. Tripathi *et al.*, "Comparison of clustering and phenotyping approaches for subclassification of type 2 diabetes and its association with remission in Indian population," *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-71126-7.
[34] J. Kaliappan *et al.*, "Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets," *Frontiers in Artificial Intelligence*, vol. 7, 2024, doi: 10.3389/frai.2024.1421751.
[35] S. Kabir, M. S. Hossain, and K. Andersson, "A review of explainable artificial intelligence from the perspectives of challenges and opportunities," *Algorithms*, vol. 18, no. 9, 2025, doi: 10.3390/a18090556.
[36] J. J. Thomas *et al.*, "Translation and impact of the national diabetes prevention program in two rural settings: participant outcomes, individual experiences, and recommendations," *Diabetology*, vol. 5, no. 7, pp. 690–705, 2024, doi: 10.3390/diabetology5070051.
[37] H. El-Sofany, S. A. El-Seoud, O. H. Karam, Y. M. A. El-Latif, and I. A. T. F. T. Eddin, "A proposed technique using machine learning for the prediction of diabetes disease through a mobile app," *International Journal of Intelligent Systems*, 2024, doi: 10.1155/2024/6688934.
[38] S. K. S. Modak and V. K. Jha, "Diabetes prediction model using machine learning techniques," *Multimedia Tools and Applications*, vol. 83, no. 13, pp. 38523–38549, 2024, doi: 10.1007/s11042-023-16745-4.

## BIOGRAPHIES OF AUTHORS

**Prajakta Bhosale-Dhamdhere** is a doctoral researcher in the Department of Computer Science and Engineering at MIT School of Computing, MIT ADT University, Pune, Maharashtra, India. Her research focuses on machine learning applications in healthcare, with specific emphasis on ensemble methods, transfer learning, and explainable AI for diabetes prediction and chronic disease management. She can be contacted at email: prajakta.dhamdhere89@gmail.com or prajaktabhosalephd2021@gmail.com.

**Dr. Ganesh Pathak** received his Ph.D. in Computer Science and Engineering, with his research focused on developing a security framework for wireless sensor networks. He is an academician and researcher with over 26+ years of experience bridging academia and industry. Currently, he serves as professor and dean, School of Computing, MIT ADT University, Pune. He has published 34 research papers published in reputed peer-reviewed journals and conference proceedings, many of which are indexed in Scopus and SCI. His research and teaching interests focus on artificial intelligence, big data analytics, and cognitive modelling. As a mentor, he guides doctoral candidates in artificial intelligence, data science, cloud computing, and security. He can be contacted at email: ganesh.pathak@mituniversity.edu.in.