# A MULTI-STAGE DIABETIC DISEASE PREDICTION MODEL USING LDA-BASED FEATURE EXTRACTION AND BGGO-ENHANCED ADAPTIVE ENSEMBLE CLASSIFICATION

**Dr.Anitha P** Assistant Professor, Department of Computer Science, Sri Ramalinga Sowdambigai College of Science and Commerce, Coimbatore

**Abstract**

The chronic and frequent asymptomatic nature of the Diabetic (DB) disease are the reason behind its major risk level globally. For prevention and therapy, the Early Detection (ED) of the DB disease is crucial. Delayed and inaccurate predictions will be result of the conventional diagnostic system, because conventional diagnostic system usually depend upon manual analysis. Issues like data imbalance, noise, and irrelevant features are the result of using current Machine Learning (ML) model. This will result in poor generalization. The prediction accuracy (ACC) is less in current methods, as it has single classifier dependency, poor feature representation, and inefficient handling of missing or imbalanced data. A comprehensive DB prediction framework integrating advanced pre-processing, clustering, feature engineering, and classification models was suggested in this study for addressing those limitations. For removing outlier noise values, pre-processing includes missing value (MV) imputation (MVI) feature encoding and feature scaling (FSc). To separate data into meaningful groups, an enhanced K-means and Fuzzy C-Means (FCM) approach is used for clustering with a consideration of uncertainty and overlapping classes. Here, feature extraction (FE) employs linear discriminant analysis (LDA) for the purpose of converting high-dimensional (HD) data into a smaller space with maximal class separability. An intelligent geese behavior was simulated by the Binary Grey Lag Goose Optimization (BGGO). For Feature Selection (FS), the BGGO is used. This BGGO assist in selecting the optimal feature subset, and maintaining classification performance. Then, Hoeffding Adaptive Tree (HAT) method is used for classification. The HAT with Synthetic Minority Oversampling Technique (SMOTE) is used, and it will gradually learn from streaming data and adjust to concept drifts. An improved Recall (R), and balancing class distribution for minority DB cases was attained by the application of the suggested method. This integrated method may offer a High-performance prediction with reduced computational complexity. The real-world healthcare data anomalies were effectively handled by this integrated method.

**Keywords:** Enhanced K-Means, Fuzzy C-Means (FCM), Linear Discriminant Analysis (LDA), Binary Grey Lag Goose Optimization (BGGO), Hoeffding Adaptive Tree (HAT), Synthetic Minority Oversampling Technique (SMOTE).

## 1. Introduction

A chronic metabolic condition with an increased blood glucose levels are termed as Diabetes Mellitus (DM). Serious risk factors related to the DM are cardiovascular diseases (CVD), kidney failure (KF), and neuropathy [1,2]. The ED and accurate prediction is needed for effectively treating disease, because the occurrence of DB is increasing globally, especially in developing nations [3]. Due to the inter-patient variability and the multi-faceted nature of DB biomarkers, the progression of the disease is not effectively detected by the conventional clinical methods [4]. For the purpose of enhancing prediction ACC, an effective prediction models (PM) using artificial intelligence (AI) and ML techniques have been analyzed. This analysis will also support the doctors in creating Decision Making (DMa) [5]. In clinical diagnostics, with the use of medical data, hidden complex patterns were revealed by the potential of ML model, but the conventional statistical models fail to detect that patterns [6]. The quality of the features and discriminability of the features obtained from the clinical datasets have a major impact on the predictive ACC [7]. Usually, the classification performance is affected by the overlapping features, but the LDA is beneficial in clinical diagnosis. For dimensionality Reduction (DR), and maintaining class-separability, this LDA has proven to be effective [8]. The most relevant clinical signs for various DB stages was identified by the LDA. Thus, the classification results across multi-stage disease progression scenarios are improved.

When dealing with imbalanced datasets and varying patient profiles, challenges in classification exist regardless of the advancements in feature extraction (FE) [9].

Multiple base classifiers (BC) predictions are combined in ensemble learning (EL) techniques. By improving generalisation and decreasing overfitting, EL provides a solid solution. Nevertheless, the efficiency of ensemble models depends on the choice of base classifiers, weight optimization, and adaptability to new data [10]. To address these limitations, metaheuristic optimization algorithms (MHOA) have been employed for tuning model parameters. However, many of these algorithms struggle with premature convergence and local optima trapping, especially in high-dimensional (HD), nonlinear (NL) spaces.

To overcome these challenges, the proposed study introduces a novel BGGO-enhanced adaptive ensemble classification approach, where BGGO stands for Binary Grey Lag Goose Optimization, a recently introduced nature-inspired algorithm (NIA) known for its strong (GS) global search and convergence capabilities. By integrating BGGO, the ensemble model dynamically adjusts classifier weights and selects optimal subsets, improving performance across multiple DB stages. High sensitivity (S) and specificity (SP), which are essential in medical diagnostics. S and SP are guaranteed by this adaptive technique. Here, the consequences of false positives (FP) and false negatives (FN) are serious

## 2. Literature Review

For DB prediction, and disease detection, the ML methods are used, and researchers also suggest a hybrid model named Diabetes Prediction Empowered with Multi-Level Data Fusion and ML (DPEMDFML). The support vector machines (SVM) with artificial (NN) neural networks (ANN) are integrated in DPEMDFML for creating a hybrid ML model. For the training and testing purpose, 2 different datasets are utilized, and this application may support in determining the model's efficiency. Train-test splits in the ratio of 70:30 and 75:25 were used for separating the dataset into 2 experimental configurations. This method offers a comprehensive analysis, and it determines the model's predictive ability. An ACC of 97.43% was attained by the ANN model, and this high score thus indicates the efficacy of the model in predicting DB. So, this suggested method is considered as a suitable and effective diagnostic tool.

A distinct deep learning (DL) method for Diabetic Foot Ulcer (DFU) identification by the utilization of convolutional neural networks (CNNs) with image processing (IP) techniques was suggested by Giridhar et al. [12]. To train the CNN model in an effective way, the DFUC 2021 dataset is used by the suggested method. The foot images that are captured using several imaging modalities are included in this dataset. Those modalities that includes infrared, thermal, and color imaging. Crucial visual features like wound size, wound type, color, and texture were analyzed, and this analysis demonstrate the robust efficiency of the trained model in the detection and classification of the DFU. For the evaluation, a multiclass (MC) (three-class) classification task is applied. From the outcomes, it is clear that this model attains remarkable values in metrics like F1 score, ACC, and Precision (P), and it also demonstrates the efficacy of the model in DFU recognition. Among the tested CNN architectures, DenseNet201 achieved the highest F1 scores, with values of 98% for Ischemia, 98% for None, and 97% for the infection stage, highlighting its robustness across different stages of DFU classification. In DB foot care scenarios, diverse and complex image data was effectively handled by the model, and it demonstrate the reliability of the model. In DB treatment, the integration of DL, especially CNN-based methods in DFU detection pipeline offers several advancements. The early diagnosis, and reducing complications are facilitated by the suggested method. This method also enhances the clinical DMa in diagnosing diabetic foot conditions.

A hybrid model named HybridFusionNet was used for the purpose of improving the classification of DB retinopathy, and this suggestion was given by Shukla et al. [13]. This HybridFusionNet model integrates the Vision Transformer (ViT) and attention mechanisms (AM). From several DR stages, the deep features are extracted, and this extraction helps the model in improving performance in both binary classification (Bc*l*) and multi-class classification. The detection ACC over both classification

tasks were enhanced by the integration of Self-Attention Network (SAN) and ViT in the HybridFusionNet. Remarkable efficiency of 91% ACC in the binary classification (BC) stage and 99% ACC in the MC stage was attained by the model. Relevant visual data from retinal images were effectively identified by the potential of the hybrid architecture, and it was revealed in the outcomes of the study. This model is suitable for obtaining accurate DR diagnosis. The integration of ViT and AM in the medical image analysis (MIA) attains remarkable efficiency over both classification tasks, and it was validated.

For the DR detection, a DiaNet Model (DNM) was suggested by Nandhini et al. [14]. In the retinal image pre-processing stage, the Gabor filter (GF) is used. The visibility of blood vessels (BV) are enhanced by this application. Activities like texture analysis, object detection (OD), FE, and image compression were facilitated by the implementation of GF. In the image augmentation (IA) stage, the input dimensions of the dataset were minimized by the application of the Principal Component Analysis (PCA). The training becomes more effective by the PCA. In specific conditions, the number of attributes are minimized by the DR, and it enhances DNM efficiency. The performance of the suggested model was compared with the current SOTA (state-of-the-art) methods for determining the efficiency of the suggested model. From the outcomes, it is clear that the suggested model outperformed current SOTA techniques with a mean classification ACC of 90.02%. The DB retinopathy is accurately detected by the DiaNet Model, and it was demonstrated by the outcomes of the analysis.

For DB classification, early-stage detection, and prediction, ML-based method was suggested by Butt et al. [15]. This study describes the hypothetical Internet of Things (IoT)-based DB monitoring system, and classification model. Here, monitoring blood glucose (BG) levels in both healthy individuals and DB patients is the main purpose of this study. For the classification task, 3 distinct classifiers are used: Random Forest (RF), Multilayer Perceptron (MLP), and Logistic Regression (LR). Then, for prediction analysis, Linear Regression (LR), Moving Averages (MA), and Long Short-Term Memory (LSTM) networks were used. The simulation is conducted by using a benchmark PIMA Indian DB dataset. During evaluation, highest classification ACC of 86.08% was attained by the MLP. The comparative analysis was performed with suggested model, and current SOTA methods. From the analysis, it is clear that the suggested method executes better than the SOTA methods in terms of DB prediction. The outcomes of the comparative analysis also demonstrate the suggested method's flexibility and resilience for possible use in actual public healthcare monitoring systems. The LSTM demonstrated superior performance, attaining an ACC of 87.26% in DB prediction.

For the ED of the DB, the several ML models were analyzed, and this analysis was conducted by Yudheksha et al. [16]. For the development process of the model, several classification methods, such as SVM, Decision Tree (DT), RF, XGBoost, K-Nearest Neighbours (KNN), and LR, were used. For hyper parameter (HP) tuning, grid search was applied for ensuring the model's optimal performance. Thus, the optimal configuration is attained by the HP, as it fine-tuned all algorithms. Standard performance metrics such as P, ACC, R, F1-score, and the ROC-AUC curve were used for the comprehensive analysis. This comprehensive analysis will help in determining the effectiveness of the suggested model. The most effective procedure for ED of DB prediction was determined by this analysis.

For DM prediction, ML-based method was introduced by Ahamed et al. [17]. The optimal 3 classifiers are then selected by the model's prediction ACC, and it was determined after a comparison and application of several ML methods. Those 3 best classifiers are: RF, Gradient (BM) Boosting Machine (GBM), and Light Gradient BM (LGBM). For experimental analysis, researchers suggest to use 2 datasets: Pima Indians Diabetes dataset and a curated dataset. With the support of the following classifiers: LGBM, GBM, and RF, the PM was created. After creating this PM, the ACC of every method has been evaluated and compared. Then, this study creates a generalized PM and the DA (Data augmentation) method for the purpose of improving the performance of the model. In 2 datasets, the impact of augmentation on prediction ACC was determined by executing the comparative analysis. From the analysis, the final prediction ACC for every classifier both with and without augmentation

 

was presented. For accurate DB prediction, the suggested ML models, specifically RF, GBM, and LGBM classifiers are effective. From the outcomes, it is also clear that the efficiency of the model was enhanced when the augmentation was applied.

A multi-model method for estimating the DB risk factors with the application of several ML methods, and it has been suggested by Jain et al. [18]. To develop the PM and estimating DB related risk factors, several algorithms, including DT, Support Vector Classifiers, Random Tree Forest, and Feedforward (NN) Neural Networks (FF) were used by the authors. By utilizing the HP tuning in the model training, the RF model executes well when configured with finite state settings and optimized estimators. The RF model has high ACC than other models involved in analysis.

The predictive efficiency was improved by the model configuration, so it highlights that the model configuration is crucial. The spearman rank correlation was used for executing statistical analysis in addition to ML-based evaluation. Thus, a significant correlation value of approximately 0.79 was attained by the risk factors such as Postprandial Glucose (PPG), Fasting Plasma Glucose (FPG), and Mean Blood Glucose (MBG), and it was revealed by the analysis. Thus, a strong interdependence value of 0.79 was attained. The strong correlation among these key factors is determined by high R-squared value, specifically among PPG and FPG. Consulting medical experts regarding the results. This will ensure clinical significance of these outcomes. For clinical DMa, the development of "dirt chart advisor" is used, and it may facilitate varied analysis and visual interpretation of risk factor relationships.

A novel hybrid ML method was suggested by Albadri et al. [19]. This hybrid model integrates the SVM, DT, and RF classifiers, and this hybrid model is designed for the purpose of improving the DB prediction. The prediction ACC was further enhanced when authors added factors to the algorithm. Those factors are age, BMI, insulin levels, and glucose levels. On the Pima Indian Diabetes dataset, the suggested hybrid algorithm was trained and validated by the application of both the holdout validation and k-fold (CV) cross-validation techniques.

From this analysis, an ACC of 88.5% on the holdout set and 90.1% with k-fold CV was attained by the hybrid model, and this suggested model executes better than the distinct classifiers. Then, 76.8% ACC was attained by the DT model, and 75.3% ACC was attained by the RF classifier. The efficacy of the model is then analyzed by using standard metrics like P, R, and F1-score. These metrics are vital, and it supports in analysing the capacity of the model in detecting true positive cases and reducing false positives (FP) cases accurately. The suggested hybrid model has proven to be an effective method for DB risk prediction and monitoring. The DB is accurately predicted by the suggested model, and it was demonstrated by the outcomes. The ED and clinical intervention is also facilitated by the suggested method.

## 3. Proposed Methodology

An enhanced adaptive ensemble classification was presented in this study for effective DB diagnosis. The DB disease is predicted in 5 stages, and it is explained in this paper. MVI is used for pre-processing in stage one. In stage two, medical data sets are evaluated using enhanced K-means and fuzzy C-means, sometimes known as K-Mean-FCM. The third procedure uses LDA to extract the pertinent features from many factors as part of the hybrid FE method. BGGO-based FS is suggested in stage four. With better results, the adaptive ensemble classification model is suggested for DB disease prediction. HAT and SMOTE are the two classifiers that are introduced for DB recognition. Figure 1 shows the suggested process.

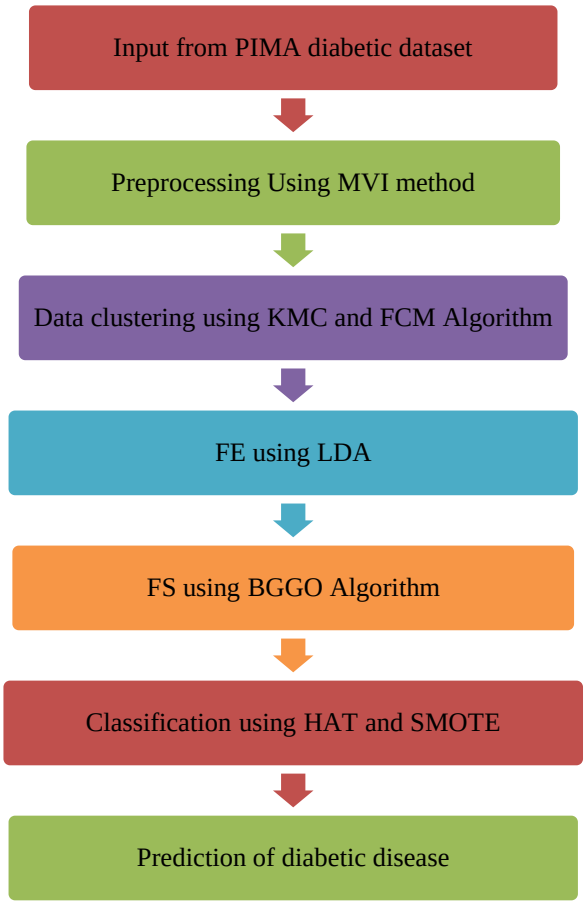


**Figure 1. A block diagram for Suggested Procedure**

### 3.1.Dataset Description

To finish DM processes, datasets are required. Numerous studies have made use of data and/or information that has already been collected from service providers, such as hospitals and other healthcare facilities. Numerous datasets are currently accessible for research. This study made use of the well-known diabetic dataset, PIDD [20]. The dataset covered 768 patient samples, including 500 non-DB and 267 DB.

### 3.2.Data Pre-Processing:

The first step involves gathering healthcare data from electronic health records (EHRs) of patients. This data includes things like demographics, vital signs, and lab results. However, this raw information might have errors, missing bits, or inconsistencies, which can mess up the predictions made by machine learning models. To address these issues, several pre-processing steps are applied:

**Data Cleaning:** The data's noise and irregularities are found and fixed. This might entail fixing typos, eliminating outliers, and fixing data format incompatibilities.

**MVI**: To deal with MV in the dataset, suitable methods such as mean, median, or k-nearest neighbours (KNN) imputation are applied. Mean imputation substitutes the average of the related feature for MV. By considering the values of similar cases in the dataset, the KNN imputation calculates MV.

$$x^{\wedge}i = \sum j - 1kxij \tag{1}$$

$$kx^{\wedge}i = k\sum j - 1kxij \tag{2}$$

For example, in a dataset containing blood glucose levels, missing values can be imputed using the mean blood glucose level of similar patients based on their demographic and clinical characteristics.

**Feature Encoding:** Categorical variables in the dataset are transformed into numerical representations so that they can be used with ML techniques. Various techniques such as label encoding and one-hot encoding can be used to accomplish this change.

$$\mathrm{xi}' = \mathrm{xi} - \min(\mathrm{x})\max(\mathrm{x}) - \min(\mathrm{x}) \quad (3)$$

$$\mathrm{xi}' = \max(\mathrm{x}) - \min(\mathrm{x})\mathrm{xi} - \min(\mathrm{x}) \quad (4)$$

A dataset containing the categorical feature "Gender" with the values "Male" and "Female," for example, would have each category converted into a binary vector via one-hot encoding, where each element would indicate whether the category is present or absent.

**FSc:** The feature values are modified to a comparable range in order to ensure consistency in feature magnitudes and avoid the dominance of features with larger values throughout the learning process. Widely-used scaling methods for this purpose include min-max scaling and standardization.

$$\mathrm{xi}' = \mathrm{xi} - \mathrm{mean}(\mathrm{x})\mathrm{std}(\mathrm{x}) \quad (5)$$

$$\mathrm{xi}' = \mathrm{std}(\mathrm{x})\mathrm{xi} - \mathrm{mean}(\mathrm{x}) \quad (6)$$

### 3.3.Data Clustering Using K-Means and FCM Algorithm

Clustering and classification are the two primary types of pattern recognition methods. Simply looping related things together is known as clustering. This method decides how to separate data into different groups based on attribute similarity. K-means algorithms and FCM algorithms are two subtypes of partitional clustering, a type of clustering. The data is simply grouped by this technique such that every object belongs to the same group.

#### K-means algorithm

A well-known clustering technique named KMC is used on the dataset in the investigation. Given that the number of operations needed to find a solution increases linearly with dataset size, one of its biggest shortcomings is how slowly it processes large datasets [37]. Here, the number of clusters is indicated by the values of $k$. Establishing $k$ centres, one for each cluster, is the fundamental idea [38]. The squared error function (SEF), an objective function (OF) that is determined as

$$\phi = \sum_{\mathrm{x}\epsilon\mathrm{X}} \min_{\mathrm{c}\epsilon\mathrm{C}} (\|\mathrm{x_i} - \mathrm{c_i}\|)^2 \quad (7)$$

Here, $\|\boldsymbol{x_i} - \boldsymbol{c_i}\|$ is the Euclidean distance between $xi\ and\ ci$. $C = \{c1, c2, c3 \ldots . cn\}$ is the number of clusters centers and $X = \{x1, x2, x3 \ldots . xn\}$ is the set of data points. An alternative name for this objective function is “distortion”

#### Elbow method

By calculating the percentage of clusters that will form an elbow at a specific location, the elbow method can be used to determine the optimal number of clusters [40]. Usually, the y-axis shows the sum of squares, and the $x$-axis shows the number of clusters. The square of the distance among every

point and its associated cluster centre is the average sum of squares following clustering. The elbow point no longer undergoes a sudden change.

**Silhouette method (SM)**

Through the use of a silhouette coefficient (SC), the SM integrates cohesion and separation. The silhouette coefficient can be calculated by taking the ratio of cohesiveness to separation and subtracting 1. A higher SC makes the cluster better. The SC is shown on the $y$-axis of the SM's graph. This approach chooses $k$ with the highest SC, and its value is shown in the $x$-axis [41]. Equations (4) and (5) describe the SC equation. The equation is as follows if the cohesion measure is higher than the separation measure (SM):

$$s = 1 - \frac{\text{seperation Measure}}{\text{cohesion measure}} \tag{8}$$

If the SM is greater than the cohesion measure, then

$$s = 1 - \frac{\text{cohesion measure}}{\text{seperation measure}} \tag{9}$$

**t-SNE plot**

By splitting high dimensional (HD) document vectors into two dimensions and using probability distributions from both the original and decomposed dimensionalities, t-SNE effectively groups related documents. A scatter plot can be used to view documents that have been divided into 2-D or 3-D.

**3.4.FE using LDA**

In this case, the varied within-class frequencies can be easily managed by LDA. Their effectiveness has been assessed using randomised test results. This strategy provides the most separability by reducing the ratio of within-class variance to between-class variation in every given batch of data. The classification difficulty in speech recognition is taken into account while classifying data using LDA. In an effort to provide a more accurate categorisation than Principal Components Analysis (PCA), researchers choose to create a method for LDA.

PCA performs more feature classification than LDA, which is the main distinction between the two techniques. Relocating the original data sets to a new location without changing their size, location, or form is what PCA recommends. LDA only aims to establish a decision area between the current classes and enhance class separability. A projection vector in the feature space that maximises the between-class scatter matrix and minimises the within-class scatter matrix becomes apparent when two-class issues are solved using the LDA method. Here, finding a linear function is the goal.

$$z = b_1 v_{j1} + b_2 v_{j2} + b_3 v_{j3} + \ldots . b_r v_{jr} \tag{10}$$

Here the vector of coefficients that needs to be identified is $b^s = [\{b_1, b_2, \ldots . , b_r\}]$.

$$v_j = \left[v_{j1}, v_{j2}, \ldots . , v_{jr}\right] \tag{11}$$

$$v_i = [v_{i1}, v_{i2}, \ldots , v_{ir}] \tag{12}$$

To estimate the mean and variance of the dataset, the following multivariate analysis of variance assumptions must be satisfied.
Normality: The independent variables at each level of the grouping variable have a normal distribution.

Collinearity: The predictive capacity may be diminished if there is a high correlation between two variables.

### 3.5.FS using BGGO

Feature selection using the sigmoid function and BGGO is a new method that enables the most relevant features for diabetes diagnosis to be found. The sigmoid function, a nonlinear (NL) activation function (AF) typically used in (NN) neural networks, is deliberately utilized to compute the relevance of the features to the target variable, which is, in this case, diabetes outcome. Large weights are attributed to those features that are believed to be more effective in predicting diabetes risk, whereas small weights correspond with those features that are less likely to be helpful.

For resolving complex problems based on the principles of the greylag geese, an optimization method that integrates the sigmoid function (SF) with BGGO are used. A version of the genetic algorithm (GA) is BGGO that doesn't simulate the natural flocking behavior of geese. For the purpose of obtaining optimal feature subset, this BGGO continuously search many feature spaces. This optimal feature subset may help in attaining the best classification performance.

It is due to the use of an evolutionary approach that the features that contribute the most to the model's accuracy are not only chosen, but also the model is stable and robust across multiple datasets and various experiments. A strong basis for FS was offered by the SF and the BGGO in a DB classifier. The feature count is reduced, and overfitting issues can be minimized by this method.

It is very promising because it improves the interpretability and generalization ability of diabetes classification models, thus providing more precise risk assessment and individual intervention strategies for diabetes patients.

$$x_d^{(t+1)} = \begin{cases} 1 \; if \; Sigmoid(m) \geq 0.5 \\ 0 \; otherwise \end{cases} \quad (13)$$

$$Sigmoid\ (m) = \frac{1}{1 + e^{-10(m-0.5)}} \quad (14)$$

**BGGO Algorithm**

BGGO is a special metaheuristic optimization algorithm dedicated to resolving complicated optimization problems, such as FS in diabetes classification. Algorithm 1, derived from population-based algorithms inspired by the flocking behavior of Greylag geese, explores solution space while identifying the optimal feature subset for classification.

**Algorithm 1 BGGO Algorithm**

*Step 1*: Initialize GGO population, OF, and GGO parameters
*Step 2:* Convert solution to binary [0 or 1]
*Step 3:* Calculate the OF for each agent and get the best agent location
*Step 4:* Update Solutions in exploration group and Exploitation group
*Step 5*: **while** $t \leq t_{max}$ **do**
*Step 6:* **for** $1 = 1\ to\ n_1$ **do**
*Step 7:* **if** $t\%2 == 0$ **then**
*Step 8:* **if** $r_3 < 0.5$ **then**
*Step 9:* **if** $|A| < 1$ **then**

*Step 10:* Update location of present search agent in exploration group
*Step 11:* **else**
*Step 12:* Update location of current search agent based on three random search agents
*Step 13:* **end if**
*Step 14*: **else**
*Step 15:* Update location of present search agent
*Step 16:* **end if**
*Step 17:* **else**
*Step 18*: Update individual locations
*Step 19:* **end if**
*Step 20:* **end for**
*Step 21:* **for** $i = 1\ to\ n_2$ **do**
*step 22:* **if** $t\%2 == 0$ **then**
*Step 23:* Update location of present search agent in exploitation group
*Step 24:* **else**
*Step 25:* Update location of current search agent
*Step 26*: **end if**
*Step 27:* **end for**
*Step 28:* Convert Updated solution to binary
*Step 29*: Calculate Objective function
*Step 30:* Update parameters
*Step 31*: Adjust beyond the search space (SS) solutions
*Step 32:* Update Solutions in exploration group and exploitation group
*Step 33*: **end while**
*Step 34:* **return** best agent

Then, the system is defined by initializing the population of BGGO agents, objective function, and parameters corresponding to the BGGO algorithm.

Converting every solution in the population into binary format, and it will represent whether the features are present or absent in the classification. After calculating an objective function for each agent, the optimal agent position is determined based on its performance. The method is iteratively explored and exploited till a termination condition is met. This termination condition is represented by the maximum number of iterations ($tmax$). Based on the iteration index ($t$) and random variables ($r3$), the condition statements that are set. Then, this method uses the condition statements for the purpose of determining the solutions of the various exploration and exploitation groups in every iteration. The algorithm can effectively use its feature space and it finally converge to an optimal feature subset, because these updates are made to strike a balance between exploration and exploitation.

### 3.6.Classification with H-SMOTE Tree

Classification using the H-SMOTE tree algorithm is the last step in the suggested method. The problem of uneven data distribution in T2DM prediction is addressed by the H-SMOTE method. It combines the HAT and SMOTE, two existing techniques. While SMOTE generates fake samples for the under-represented class to ensure the model isn't biassed towards the more prevalent conclusion, HAT assists the model in adapting to new data as it becomes available.

**HAT:**

A decision tree-based classifier called HAT gradually constructs the decision tree structure while gradually adjusting to shifts in the distribution of the input. When dividing the feature space according to the purity of the class labels, the Gini index is frequently employed as the splitting criteria.

$$\text{GiniIndex} = 1 - \sum \text{i} = 1\text{kpi2} \qquad (15)$$

For instance, the Gini index is computed for every feature in a dataset containing several features in order to identify the best split that maximises the distance between patients with T2DM and non-T2DM.

**SMOTE:**

By generating additional data points (DP) for the less prevalent group (minority class), SMOTE aids in data balance. It accomplishes this by selecting pre-existing DP from that group and generating new ones that are in the middle. By doing this, the class distribution is balanced and the model is not skewed towards the more typical result.

$$xnew = xi + \lambda \times (xj - xi) \quad (16)$$

Where: $\boldsymbol{xi}$ and $\boldsymbol{xj}$ are instances of the minority class. $\boldsymbol{\lambda}$ is a random value between 0 and 1. Within the H-SMOTE tree algorithm, synthetic samples are crafted for the minority class as part of the training phase, guaranteeing that the classifier is trained on a dataset with balanced class representation. The structure of the DT is dynamically modified according to the attributes of the input data, enabling the model to accommodate variations in class distribution as they occur. During classification, the H-SMOTE tree algorithm assigns class labels to new instances based on the majority class of the corresponding leaf node in the decision tree.

## 4. RESULT AND DISCUSSION

DB ED aims to improve wellbeing and increase life expectancy. A variety of supervised and unsupervised algorithms have been proposed for the prediction of diabetes in its early stages. This section looks at the dependability of LDA-FE and shows promising results for several variables. The evaluation was conducted using MATLAB. Here, researchers provide a few simple experiments in which we used a particular dataset and the suggested BGGO-based architecture. Performance metrics like sensitivity (S), specificity (SP), and precision (P) coefficient are then computed using the following formulas:

P is determined by dividing the number of correctly identified positive observations by the total number of expected positive observations.

$$P = TP/(TP+FP) \quad (17)$$

The S ratio (or) Recall (R) is the ratio of correctly detected positive observations to all observations.

$$R = TP/(TP+FN) \quad (45)$$

The F1 score is the weighted average of ACC and R. Thus, FP and FN are acceptable. ACC is computed using positive and negative values in the following formula.

$$F1\ Score = 2*(R * P) / (R + P) \quad (18)$$

The following is how ACC is determined in terms of positives and negatives:

$$ACC = (TP+FP)/(TP+TN+FP+FN) \quad (19)$$

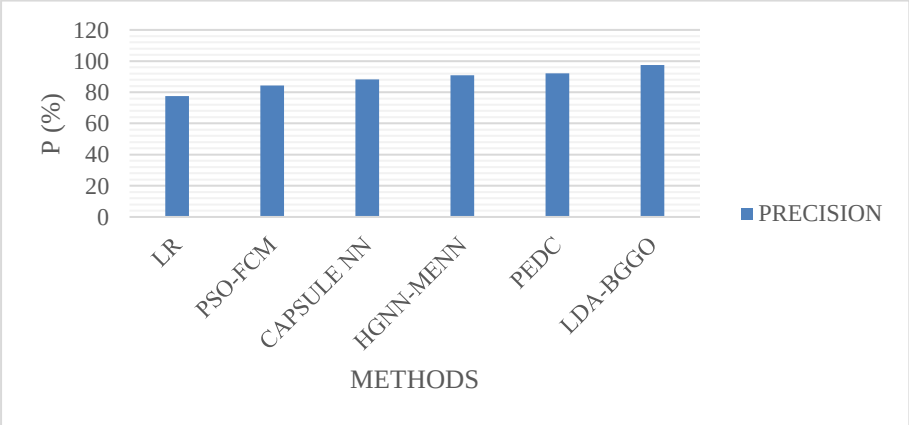


Figure.2. P Comparison Of The Suggested and Current Technique

Figure 2 compares the precision (%) of various methods used in a classification task. The methods evaluated include LR, PSO-FCM, Capsule NN, HGNN-MENN, PEDC, and LDA-BGGO. Among these, LR (Logistic Regression) shows the lowest precision, under 80%, indicating relatively weaker performance. Precision improves progressively across PSO-FCM, Capsule NN, HGNN-MENN, and PEDC, all of which achieve precision values above 85%. The LDA-BGGO method demonstrates the highest precision, nearing or slightly surpassing 95%, highlighting its superior performance in accurate classification. This analysis suggests that advanced hybrid and deep learning-based methods significantly outperform traditional and standalone models in terms of precision.

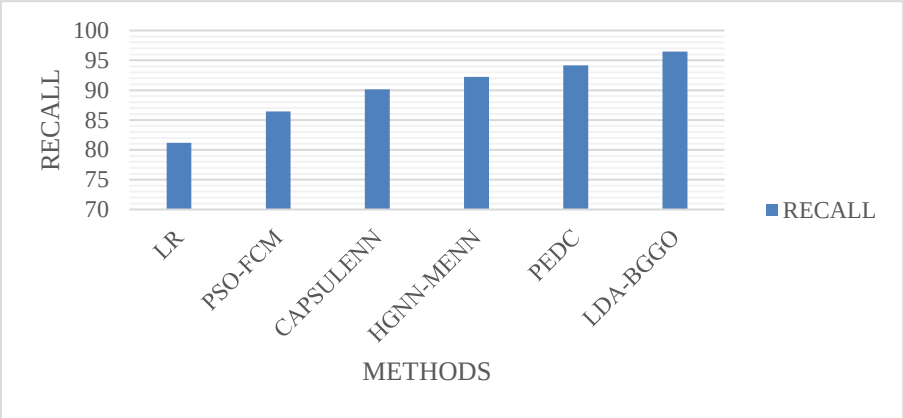


Figure. 3. R Comparison of the Suggested and current approaches

Figure 3, demonstrates the recall performance of six different classification methods: LR, PSO-FCM, Capsule NN, HGNN-MENN, PEDC, and LDA-BGGO. Among these, LR (Logistic Regression) has the lowest recall, around 81%, indicating a higher rate of false negatives. Performance improves with PSO-FCM and Capsule NN, which reach approximately 86% and 90% recall, respectively. More advanced techniques such as HGNN-MENN and PEDC show continued improvement, with recall values around 92% and 94%. Notably, LDA-BGGO achieves the highest recall, surpassing 96%, suggesting that it is the most effective at correctly identifying relevant instances. This trend demonstrates that hybrid and deep learning-based methods offer significantly better recall capabilities compared to traditional models.

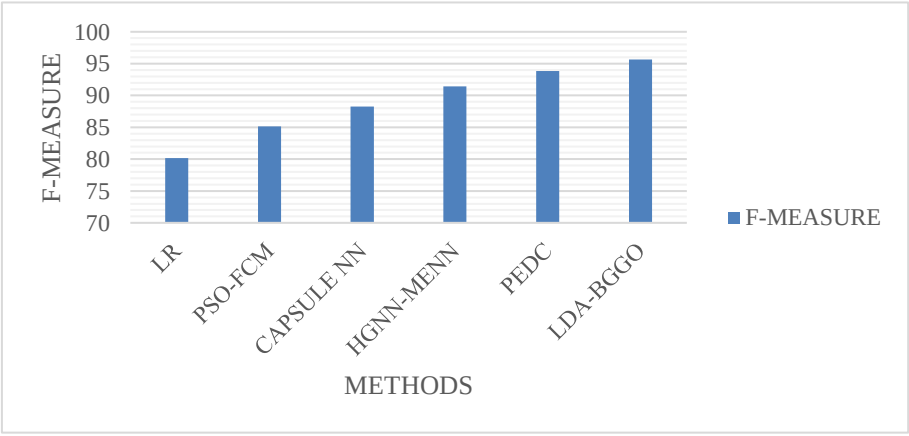


Figure.4. F-Measure Comparison of the Suggested and current approaches

Figure 4, illustrates the F-measure performance of six classification methods, showing a clear progression in effectiveness. Traditional LR achieves the lowest score (~80), while hybrid and advanced models like PSO-FCM, Capsule NN, and HGNN-MENN show steady improvement. PEDC and LDA-BGGO outperform all others, with LDA-BGGO achieving the highest F-measure (~96), indicating superior precision-recall balance. Overall, the graph highlights the enhanced classification accuracy of advanced techniques over conventional models.

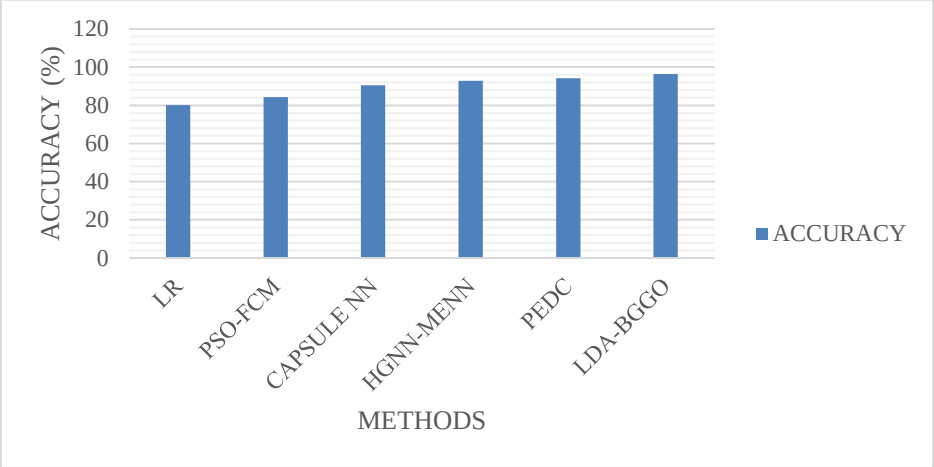


Figure.5 ACC Comparison of the Suggested and current approaches

Figure 5, presents the accuracy (%) of six classification methods, revealing a consistent improvement across models. LR records the lowest accuracy (~80%), followed by PSO-FCM and Capsule NN, indicating moderate performance. More advanced models like HGNN-MENN, PEDC, and LDA-BGGO show notably higher accuracies, with LDA-BGGO achieving the highest (~96%). This trend underscores the superior predictive performance of hybrid and deep learning-based techniques compared to traditional approaches.

## CONCLUSION

In conclusion, the proposed multi-stage diabetic disease prediction model effectively integrates LDA-based feature extraction with a BGGO-enhanced adaptive ensemble classification strategy to achieve high accuracy and robustness in early diagnosis. By leveraging Linear Discriminant Analysis (LDA), the model efficiently reduces dimensionality while preserving class-discriminative information, and the Binary Greylag Goose Optimization Algorithm (BGGO) algorithm enhances ensemble classifier performance by optimally tuning classifier parameters and selection. This synergy significantly improves prediction accuracy, sensitivity, and specificity across different stages of diabetes. For future work, the model can be extended by incorporating longitudinal patient data, exploring deep learning-based hybrid feature extraction techniques, and deploying the framework in real-time clinical decision support systems to evaluate its practical applicability and scalability across diverse healthcare settings.

## REFERENCES

1. Rodriguez, K., Ryan, D., Dickinson, J.K. and Phan, V., (2022). Improving quality outcomes: the value of diabetes care and education specialists. *Clinical Diabetes*, vol.*40*, no.3, pp.356-365.
2. Doyle-Delgado, K. and Chamberlain, J.J., (2020). Use of diabetes-related applications and digital health tools by people with diabetes and their health care providers. *Clinical Diabetes*, vol.*38*, no.5, pp.449-461.
3. Choudhury, A., & Gupta, D. (2019). Machine learning approach for detection of diabetes. Procedia Computer Science, vol.165, pp.88–95.
4. Verma, S. and Rattan, P., (2021). Introduction to Data Mining Tools and Techniques & Applications: a Review. *in Business*, p.57.
5. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, vol.3, pp.1157–1182.
6. Jia, L., Ma, X., & Wang, X. (2020). An effective LDA-based feature selection method for sentiment classification. Pattern Recognition Letters, vol.136, pp.142–149.
7. Dietterich, T. G. (2000). Ensemble methods in machine learning. In International Workshop on Multiple Classifier Systems pp. 1–15.
8. Zhou, Z.H., (2025). Ensemble methods: foundations and algorithms. CRC press.

9. Xue, B., Zhang, M., & Browne, W. N. (2013). Particle swarm optimization for feature selection in classification: A multi-objective approach. IEEE Transactions on Cybernetics, vol.43, no.6, pp.1656–1671.
10. Wang, H., & Zhao, W. (2022). Binary Grey Lag Goose Optimization (BGGO) for feature selection in classification problems. Expert Systems with Applications, vol.193, pp.116460.
11. Bassam, G., Rouai, A., Ahmad, R. and Khan, M.A., (2023). Diabetes prediction empowered with multi-level data fusion and machine learning. *International Journal of Advanced Computer Science and Applications*, vol.*14*, no.10.
12. Giridhar, C., Akhila, B., Kumar, S.P. and Sumalata, G.L., (2024, June). Detection of Multi Stage Diabetes Foot Ulcer using Deep Learning Techniques. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pp. 553-560.
13. Shukla, A., Tiwari, S. and Jain, A., (2024). HybridFusionNet: Deep Learning for Multi-Stage Diabetic Retinopathy Detection. *Technologies*, vol.*12*, no.12, p.256.
14. Nandhini, S., Sowbarnikkaa, S., Mageshwari, J. and Saraswathy, C., (2023, May). An automated detection and multi-stage classification of diabetic retinopathy using convolutional neural networks. In *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, pp. 1-5.
15. Butt, U.M., Letchmunan, S., Ali, M., Hassan, F.H., Baqir, A. and Sherazi, H.H.R., (2021). Machine learning based diabetes classification and prediction for healthcare applications. *Journal of healthcare engineering*, vol.*2021*, no.1, p.9930985.
16. Yudheksha, G.K., Murugadoss, V., Reddy, P.S., Harshavardan, T. and Sriramulu, S., (2022, December). A Machine Learning based Approach to Detect Early Stage Diabetes Prediction. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, pp. 919-924.
17. Ahamed, B.S., Arya, M.S. and Nancy, A.O.V., (2022). Diabetes mellitus disease prediction using machine learning classifiers with oversampling and feature augmentation. *Advances in Human-Computer Interaction*, vol.*2022*, no.1, p.9220560.
18. Jain, R., Tripathi, N.K., Pant, M., Anutariya, C. and Silpasuwanchai, C., (2024). Investigating gender and age variability in diabetes prediction: a multi-model ensemble learning approach. *IEEE Access*.
19. Albadri, R.F., Awad, S.M., Hameed, A.S., Mandeel, T.H. and Jabbar, R.A., (2024). A Diabetes Prediction Model Using Hybrid Machine Learning Algorithm. *Mathematical Modelling of Engineering Problems*, vol.*11*, no.8.
20. Sohail, M. N., Jiadong, R., Muhammad, M. U., Chauhdary, S. T., Arshad, J., &Verghese, A. J. (2019). An accurate clinical implication assessment for diabetes mellitus prevalence based on a study from Nigeria. *Processes*, vol.*7*, no.5, pp. 289.