International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(5s) 2026 pp. 133-153
https://doi.org/10.70917/ijcisim-2026-2681
Article

An Explainable Lightweight Deep Learning Framework for Early Diabetes Risk Prediction Using Clinical Dataset and Hybrid Optimization Techniques

S. Sai Prakash[1]*, A. C. Subhajini[2]

[1] Department of Computer Applications, Noorul Islam Centre for Higher Education, Thuckalay, Kumaracoil – 629180, Tamil Nadu, India.
Email: saiprakash.niche@gmail.com
ORCID: 0009-0009-4152-4218
[2] Department of Computer Applications, Noorul Islam Centre for Higher Education, Thuckalay, Kumaracoil – 629180, Tamil Nadu, India.
Email: jinijeslin@gmail.com

Abstract: Early diabetes risk identification is vital for preventing progression to disease and managing long-term complications of health care. However, many existing AI risk prediction models suffer from poor generalization due to several issues; class imbalance, irrelevant clinical features, lack of explainability, high computational complexity, and inadequate validation on heterogeneous patient data sets. This research examines these issues utilizing the Early Stage Diabetes Risk Prediction Dataset 2025, which contains demographic, clinical, and symptom attributes that can be used for early diabetes risk screening purposes. The framework includes missing value imputation, z-score normalization, data augmentation using Conditional Tabular Generative Adversarial Network (CTGANs), mutual information and chi-square feature selection, hybrid optimization methods, and Integrated Gradient-based methods for providing explanation of predictions. Classification is performed using a low computational complexity convolutional model, Lightweight Convolutional Integrated Gradients-Griffon Vultures Optimization Algorithm and Revolution Optimization Algorithm (LwCIG-GVROA). Experimental evaluation resulted in 95.20% accuracy, 96.22% precision, 97.12% recall and 99.65% F1 score; indicating that this approach to early diabetes risk prediction has good reliability, computational efficiency, and clinical interpretability.

Keywords: Early diabetics risk prediction, Hybrid deep learning optimization, Feature selection, Z-score normalization, Chi-square tests.

1. Introduction

Diabetes is a serious and ongoing metabolic disorder that affects over 500 million individuals worldwide (Stoleru et al. 2024). When a person experiences elevated glucose levels, it can be due to either a lack of insulin produced by the pancreas or to the body not using insulin as efficiently as it should (Ahmed et al. 2024). Individuals with early-stage diabetes often do not exhibit clear or obvious symptoms that will assist in their detection; therefore, targeted screening can be utilized to identify individuals who are at risk of the disease at an early stage before developing complications such as cardiovascular illness, kidney disease & failure, neuropathy, etc (Jadon et al. 2024). The sudden loss of weight or feeling fatigued can be leveraged with patient history data to predict the likelihood of developing diabetes at an early stage. By leveraging these clinical signs, it is possible to stratify patients by risk level (Pokhrel et al. 2025). The early recognition and management of diabetes have become exceptionally important for controlling morbidity and progression related to diabetes and its consequences (Khunti et al. 2025).

Combining MI & AI models enables more rapid and accurate predictions by leveraging patterns across large health databases to deliver proactive preventative intervention before the development of severe symptoms (Hossain et al. 2024). Access to diagnostic tools and treatment protocols enables management and regulation across every region of the world (Ojurongbe et al. 2024). Complications from diabetes will include nephropathy, cardiovascular disease, and retinopathy (Liu et al. 2025). Due to poor glycemic management, diabetic patient mortality and readmission rates are still high (Zeinalnezhad and Shishehchi 2024). By predicting diabetes at an early stage, patients at risk can be identified, enabling early intervention. By increasing early detection, both patient outcomes and healthcare costs associated with disease progression can be reduced (Olabanjo et al. 2025).

With the rise in diabetes, artificial intelligence (AI) has become a useful tool for predicting and managing the disease early. In addition, AI technologies are being used for risk assessment of complications (Teixeira et al. 2024). Algorithms in MI are helping to identify and create preventative measures and focused treatment options for the increasing number of patients who are at a higher risk of developing type 2 diabetes. Algorithms can incorporate clinical, lifestyle, and genetic variables to improve prediction accuracy (Yadalam et al. 2024). By alerting healthcare teams early and utilizing early detection, providers can provide preventive care and decrease readmission rates. The use of AI in diabetes enables enhanced health system delivery, improved health outcomes, and an overall strengthening of health systems (Gowthami et al. 2024). AI is adopted in healthcare by considering ethical responsibility, real-world implementation, and data governance. Early disease detection models ensure transparency in decision-making, support healthcare data protection regulations, and reduce algorithmic bias to maintain patient trust. Thus, solving this concern is important to translate AI-based risk prediction systems from experimental settings into routine clinical practice.

In health informatics, predicting risk of early diabetes transforms raw patient data into actionable clinical insights. In addition, AI-based risk prediction models that integrate with clinical workflows assist nurses, public health professionals, and physicians in prioritizing high-risk individuals for lifestyle counseling, screening, and follow-up care (El-Sofany et al. 2024). From the perspective of social care contexts, predictive systems help community health workers identify vulnerable populations and implement preventive programs in low-resource and rural settings. Hence, interpretable and computationally efficient methods are suitable for trust, usability, and real-world clinical adoption, in addition to predictive accuracy (Khalifa and Albadawy 2024).

Early prediction of diabetes risk intersects clinical medicine, policy health policy, behavioral health, and organization for healthcare. In behavioral aspect, interpretable and transparent risk predictions can motivate patients to engage in physical activity, adopt healthier dietary habits, and follow preventive care recommendations. Moreover, the health strategies of population can be implemented using AI-based screening tools at the policy level that enable efficient healthcare resources and targeted interventions. In terms of organization, the lightweight and explainable predictive systems are important for clinicians, community health workers and nurses. Hence, the effective prediction of diabetes in the early stage addresses policy relevance, organizational feasibility and behavioral adoption.

The motivation behind this research is to create models that are accurate, interpretable, and computationally efficient for detecting diabetes at an early stage. The MI-based methods currently in use do not excel at handling data imbalance between classes, noisy or relevant features, or low transparency and high resource requirements, which affect their viability and suitability for clinical use. Consequently, the proposed LwCIG-GVROA framework is being developed to provide superior robustness and transparency to existing MI-based systems by combining LwCNNs and XAI techniques with hybrid optimization algorithms to produce more robust, transparent, and generalizable predictions. This new approach will facilitate prompt intervention into patient care, leading to improved outcomes.

The major contributions are as follows:

- To develop a precise, understandable model for forecasting early-stage diabetes risk. To tackle data imbalance, concentrate, and interpretability issues. To focus on using the framework with lightweight/optimized AI models to improve clinical implementability.
- The newly created mixed framework, LwCIG_GVROA, merges a Lightweight Convolutional Neural Network (LwCNN) with IG for interpretable predictions and will be optimized using a hybrid GVROA to achieve efficient convergence and high-performance predictions.
- Missing data will be handled via z-score normalization and regression-based imputation, whilst MI and Chi-square testing will increase feature importance. Synthetic data generation will use a CTGAN to address class imbalance, thereby enhancing the model's overall robustness.
- The proposed framework achieves high predictive performance with low computational cost. IG improves transparency to identify clinically significant risk factors that support patient counseling and

clinical decision-making by enabling explainable predictions within social care workflows and health informatics.

The remaining part follows a similar structure: Section 2 presents the literature review; Section 3 discusses early diabetic risk prediction; Section 4 presents the findings and offers a discussion; and Section 5 offers a conclusion.

2. Literature survey

(Doğru et al. 2023) proposed a new super ensemble learning model that uses a meta-learner and four base learners. The datasets for diabetes in 130 US hospitals (98%) and PIMA (92%), as well as early-stage diabetes risk prediction (99.6%), achieved the highest model accuracy. High statistical scores and the best feature selection method, chi-square, indicated model's resilience.

(Rastogi and Bansal 2023) introduced to enhance the accuracy of early diagnosis accuracy. Their model integrates feature selection and classification methods to identify key diabetes indicators from medical datasets. The approach aims to improve predictive performance and assist healthcare professionals in making timely decisions.

(Gündoğdu 2023) developed a successful strategy for predicting diabetes in its early stages by combining the Random Forest-based feature selection method with the XGBoost classifier. To reduce dimensionality and improve model accuracy of the Random Forest algorithm by identifying relevant features. The relevant features selected by the Random Forest algorithm to make predictions for the XGBoost classifier at a high level of performance.

(Zhou et al. 2023) investigated that it is possible to build a diabetes prediction model using the Boruta feature selection method to discover the most relevant features from medical records. Next, predictive accuracy was improved through an ensemble of various classifiers using ensemble learning, providing increased robustness and overall predictive capability for diagnosing diabetes.

(Haldorai et al. 2024) focused on an AI-based model to identify maternal health risks through an analysis of clinical and demographic data to determine levels of risk using MI algorithms and to determine the most important factors that will influence maternal outcomes. This is a model that supports early diagnosis and decision-making, providing better healthcare interventions to improve maternal outcomes.

(Oliullah et al. 2024) focused on a stacked ensemble MI model consisting of random forest, decision tree, and gradient boosting as the base learners to develop a more accurate diabetes risk prediction. There is also a strong emphasis on using feature selection techniques to improve model performance by reducing overfitting, as well as on significant validation against actual clinical data, demonstrating improved prediction accuracy compared to the base models.

(Sun et al. 2025) A framework was developed using MI to evaluate diabetes risk from electronic patient records. It tests various ML techniques on their ability to predict diabetes. Its goal is to provide the framework to facilitate early diagnosis and personalized healthcare interventions. A comparison with previous work can be presented in Table 1 of this article.

Table 1: Comparative analysis of the recent work

Author	Method	Achievement	Limitations
Dogru, A., et al. (2023)	Hybrid Super Ensemble Learning	Achieved high prediction accuracy in early-stage diabetes detection using a hybrid ensemble model	Dataset details and real world validation lacking
Rastogi, R. and Bansal, M., et al. (2023)	Data Mining Techniques	Developed a model for diabetes prediction using classical data mining techniques	Specific algorithms, dataset, and performance metrics not elaborated
Gundogdu, S., (2023)	XGBoost	Efficient early-stage diabetes prediction with improved feature selection	Limited to specific classifiers and may not generalize well to larger datasets

Zhou, H., et al. (2023)	Ensemble Learning	Enhanced diabetes prediction performance using Boruta for robust feature selection	Dataset details and model interpretability not addressed
Haldorai, A., et al. (2024)	AI-based Maternal Risk Prediction Model	Predicts maternal health risks using AI methods	Not directly focused on diabetes; lacks diabetes-specific modeling
Oliullah, K., et al. (2024)	Stacked Ensemble ML	Improved diabetes prediction accuracy through stacking multiple models	Complexity in model integration and interpretation
Sun, Q., et al. (2025)	ML-based Risk Assessment Model	Real-world health data	Need for deeper clinical validation and cross-regional dataset testing.

2.1 *Problem Statement*

Most studies have provided incomplete information about their datasets (i.e., demographics, sample sizes, and data quality), making it difficult to determine how well any given model generalizes. Additionally, although some algorithms and performance metrics are discussed, gaps remain in evaluating predictive accuracy and reliability. Additionally, existing models are primarily limited to a small subset of classifiers and do not adapt well to larger or more diverse data sets. Model interpretability is frequently overlooked and is an essential component of clinical adoption and decision-making. Most approaches do not specifically address diabetes risk factors; consequently, they are not highly relevant in making targeted predictions. The complexity of establishing multiple models creates challenges in implementing and interpreting the results. Also, there is a strong need for clinical validation and testing of models in other areas to achieve broader of applicability. The current validation and methodology are unlikely to support effective early identification of diabetes. Addressing issues related to the reliability, interpretability, and general clinical use of the early prediction system for diabetes is crucial to achieving successful early identification of diabetes.

3. Proposed methodology

The proposed strategy presents a diabetes risk detection algorithm that preprocesses the collected data using techniques such as z-score normalization and missing-value imputation. Feature selection methods, such as the MI and chi-square tests, are then used to reduce dimensionality while preserving relevant information. To provide better-quality data for learning, data augmentation techniques such as CTGAN are utilized to enlarge the original dataset. Subsequently, LwCNN & IG are utilized for diabetes risk detection, and the LwCIG algorithm is trained using a hybrid procedure combining the Griffon Vulture Optimization Algorithm (GVOA) and the Revolution Optimization Algorithm (ROA), ultimately leading to improved model prediction accuracy. This hybrid GVOA-ROA approach improves convergence speed and parameter tuning efficiency. Furthermore, IG is used for model interpretability, highlighting the contribution of individual features in the prediction outcome. Figure 1 shows the early risk prediction for diabetes.

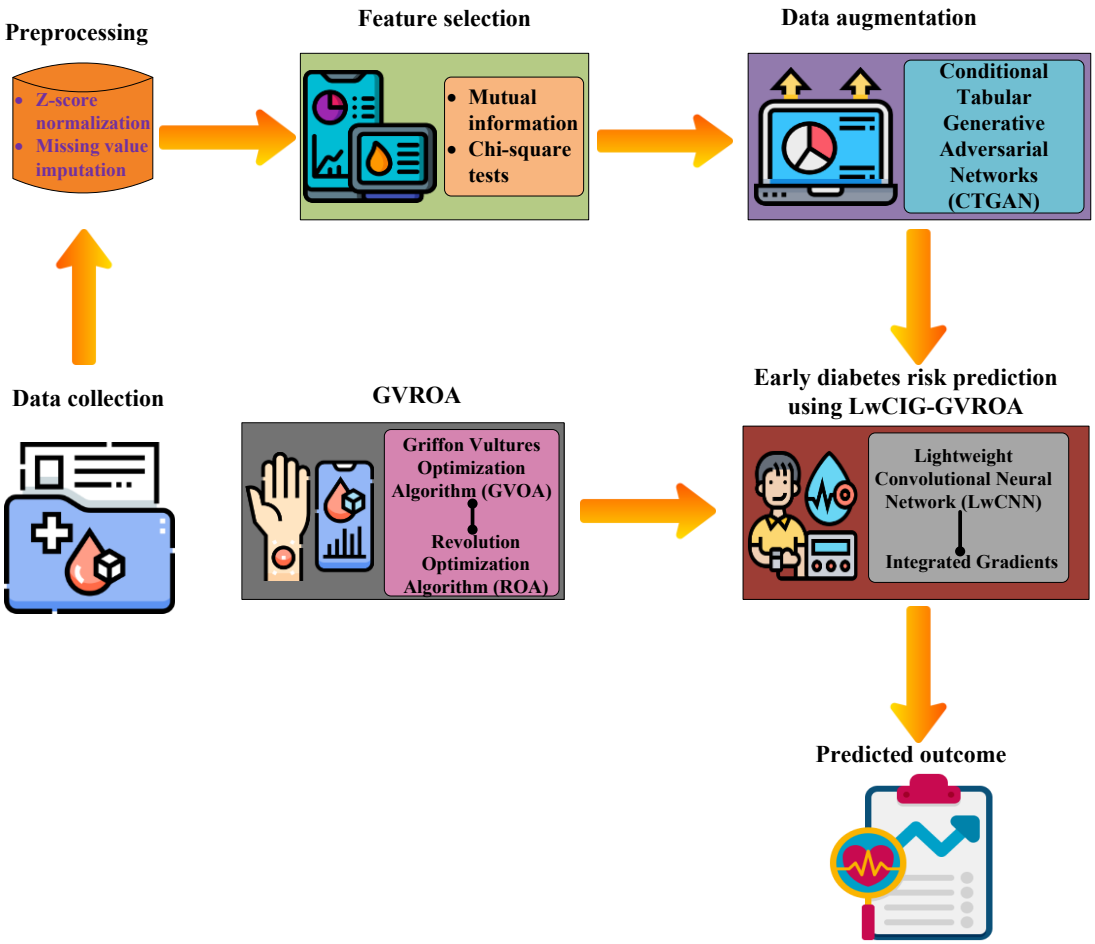

Figure 1: Proposed method for early diabetics risk prediction

3.1 Data collection

The Early Stage Diabetes Risk Prediction Dataset 2025 is health-related data collected from patients, primarily focused on clinical features useful for predicting early signs of diabetes. It includes attributes such as age, polydipsia, gender, weakness, sudden weight loss, and other relevant symptoms. The patient is at risk of developing diabetes, enabling supervised ML models to classify new instances effectively. This dataset is particularly useful for building early diagnostic tools to identify diabetes risk before it progresses. Assume the dataset with the quantity of stage diabetes risk prediction, which can be given by

$$J = \{S_1, S_2, ...S_j, S_n\} \quad (1)$$

In Equation (1), denotes the count of data, and represent the number or data. The dataset is subjected towards preprocessing.

3.2 Data preprocessing

Data is preprocessed using handling missing values by imputation techniques. Z-score normalization is applied to standardize the feature scales.

Z-score Normalization

It (Xu et al. 2025), also known as standardization, modifies data by subtracting the dataset's mean and dividing by its standard deviation. As a result, the distribution has a standard deviation of one and a mean of zero. It is especially useful when features have different scales or units, ensuring fair comparison across variables. The features are normalized by subtracting the mean values initially and then dividing the deviations, as shown in Equation (2),

$$A^* = \frac{A - \overline{A}}{\sigma A} \quad (2)$$

Z-score normalization is applied when features are transformed to have a mean of 0 and a standard deviation of 1. When the neural network learns the mapping between normalized input and output features, the mapping between original input and output features is restored by de-normalization.

$$A = A^* \Theta \sigma \mathrm{A} + \overline{\mathrm{A}} \quad (3)$$

Equation (3) shows Θ wise product of the element. It is the most commonly used methods and shared the mathematical foundation of the same with other normalization techniques.

Missing value imputation

The mean and median are commonly used to impute missing data in diabetes prediction, a process known as missing-value imputation (Olisah et al. 2022). However, these methods sometimes greatly increase the likelihood of data bias. Another method is to use multiple imputations of missing values, such as Multiple Imputation by Chained Equations (MICE). The MICE method is known to outperform the mean and median techniques, though it performs worse when the predictor variables are nonlinear. Missing-value imputation is performed predictively using Polynomial Regression (PR), a nonlinear regressor. For feature extraction, the preprocessed output is placed through the MI and Chi-Square tests.

3.3 Feature selection using MI and Chi-Square tests

Data is preprocessed using handling missing values by imputation techniques. Z-score normalization is applied to standardize the feature scales.

Z-score Normalization

It (Xu et al. 2025), also known as standardization, modifies data by subtracting the dataset's mean and dividing by its standard deviation. As a result, the distribution has a standard deviation of one and a mean of zero. It is especially useful when features have different scales or units, ensuring fair comparison across variables. The features are normalized by subtracting the mean values initially and then dividing the deviations, as shown in Equation (2),

$$A^* = \frac{A - \overline{A}}{\sigma A} \quad (2)$$

Z-score normalization is applied when features are transformed to have a mean of 0 and a standard deviation of 1. When the neural network learns the mapping between normalized input and output features, the mapping between original input and output features is restored by de-normalization.

$$A = A^* \Theta \sigma \mathrm{A} + \overline{\mathrm{A}} \quad (3)$$

Equation (3) shows Θ wise product of the element. It is the most commonly used methods and shared the mathematical foundation of the same with other normalization techniques.

Missing value imputation

The mean and median are commonly used to impute missing data in diabetes prediction, a process known as missing-value imputation (Olisah et al. 2022). However, these methods sometimes greatly increase the likelihood of data bias. Another method is to use multiple imputations of missing values, such as Multiple Imputation by Chained Equations (MICE). The MICE method is known to outperform the mean and median techniques, though it performs worse when the predictor variables are nonlinear. Missing-value imputation is performed predictively using Polynomial Regression (PR), a nonlinear regressor. For feature extraction, the preprocessed output is placed through the MI and Chi-Square tests.

3.3 Feature selection using MI and Chi-Square tests

Feature selection is used for MI (Yu et al. 2025) and Chi-Square tests to identify the most relevant features for a classification problem of target variables. MI measures the amount of information that one variable provides, capturing both linear and non-linear relationships. Chi-Square tests evaluate whether a significant association exists between two categorical variables by comparing observed and expected frequencies.

MI

MI is used to quantify the relationship between the amount of information and random variables. In this case, the one variable regarding the other can be treated as a single variables. To measure the level of dependence between MI features B and the label C, including nonlinear relationships. If the two vectors $MI(B;C)$ between the vectors are B and C is 0, then these two variables do not give any information regarding each other. Equation (3) shows the MI formula,

$$MI(B;C) = \sum_{b \in B} \sum_{c \in C} e(b,c) \log\left(\frac{e(b,c)}{e(b)e(c)}\right) \tag{4}$$

In Equation (4), MI combined with entropy, it is denoted as the entropies of B and C. The $e(b,c)$ represent joint probability distribution of B and C. When $e(b)$, $e(c)$ denotes marginal probabilities of B and C. The extracted features using MI can be denoted as f_1.

Chi-Square tests

Chi-Square tests (Ahakonye et al., 2023) are used to assess the independence between each target variable and each feature, especially in classification problems. It measures the observed frequency distribution of categorical data relative to the expected distribution. Features with high Chi-Square are more dependent on the target and thus more informative. The term B^2 evaluate the strength of association between a feature and the target class based on the observed frequencies and expected, it is described in Equation (5)

$$B^2 = \sum_{j=1}^{l} \frac{\left(P_j - F_j\right)^2}{F_j} \tag{5}$$

where, B^2 denotes the association of categorical features to classes. Feature selection from the data is denoted as $M_c \cdot$ The extracted features using Chi-Square tests can be denoted as f_2. The feature vector of both feature output can be,

$$F = (f_1, f_2) \tag{6}$$

Then, the extracted feature vector F is passed towards CTGAN for data augmentation.

3.4 Data augmentation using CTGAN

CTGAN (Khosravi et al. 2024) is used to generate synthetic tabular data for data augmentation. They are designed to handle both categorical and numerical features effectively. The specific values of CTGAN capture complex patterns and relationships in the data. It is useful for addressing class imbalance and improving the model training. Classifier performance on the minority positive risk class can be improved by using CTGAN to generate high-quality synthetic tabular data while preserving inter-feature interactions and class-conditional distributions. The generated data is real, enhancing the robustness of ML models.

The CTGAN method models tabular data distributions and generates synthetic data using Generative Adversarial Networks (GANs). The CTGAN employs a training-by-sampling technique with a conditional generator to handle non-Gaussian and multimodal distributions. To deliver outstanding results, the model design incorporates several cutting-edge processes. While preserving the fundamental learning methodology of GANs, CTGAN uses a specific set of criteria to improve the data-generating process. In a competitive training method, two neural networks,

a Discriminator (D) and a Generator (G) are used to produce data that resembles real samples. At the same time, D tries to differentiate between synthetic and real samples. Figure 2 exhibits CTGAN synthetic data.

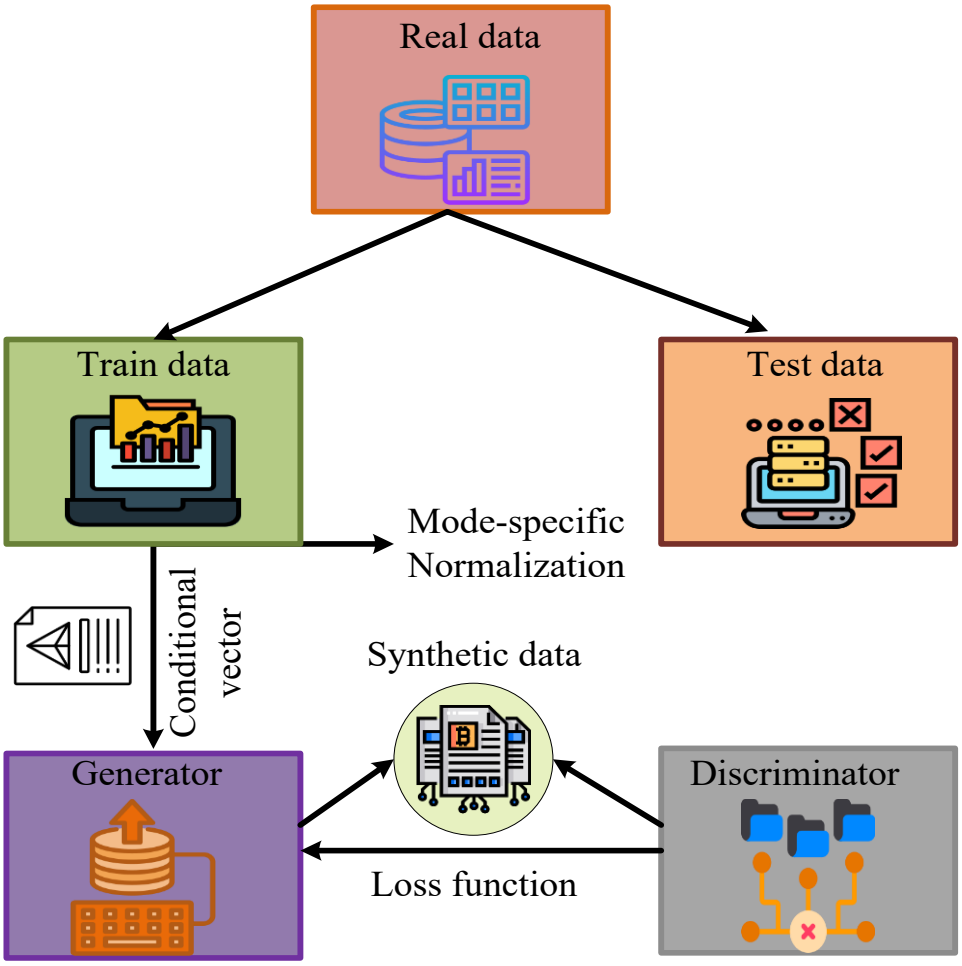

Figure 2: CTGAN to generate synthetic data

Data augmentation output C_t is subjected to early diabetics risk prediction using LwCIG.

3.5 Early diabetics risk prediction using LwCIG

Within the LwCIG algorithm, diabetes risk is determined by combining IG (Sheth et al., 2024) and LwCNN to produce a compact yet interpretable model that is computationally efficient, exhibits a hierarchical structure, and captures hierarchical relationships in structured patient information. The LwCNN is a deep learning model designed to minimize complexity while maintaining high predictive power to solve early diagnosis cases. Adding IG to this model makes it more transparent by providing an attribution for each input feature, thereby enabling understanding of how to make medical decisions.

LwCNN

The LwCNN is a deep learning model that has been simplified to be efficient on devices with minimal computational resources, such as embedded systems. They also reduce the number of parameters and operations by using methods such as quantization, depth-wise separable convolution, and pruning, thereby reducing the amount of storage needed to train a model. A dropout layer is included to help prevent overfitting, and after each convolution layer, a leaky ReLU activation function is used to allow the network to learn complex patterns. A ReLU function outputs 0 when the input is less than 0 and outputs the same value as the input when the value is greater than or equal to 0. Instead of using a flat slope for negative values, the Leaky Rectified Linear Unit, which is developed from ReLU, uses an offset slope. The network can learn from negative values because of its architecture, which keeps neurons

from being completely deactivated. This improvement increases the network's ability to recognize complex patterns in the data. The following Equation (7) is defined by:

$$h(y) = \max(0, y) \tag{7}$$

In order to obtain the output appropriate for fully connected layers, it was flattened. The term M_2 regularization sometimes referred to as Ridge regularization, and the λ is the regularization strength or hyper parameter, which regulates the trade-off between fitting the training data and minimizing the size of the weights, whereas w_j stands for an individual weight in the neural network.

$$M_2(w) = \lambda \times \sum w_j^2 \tag{8}$$

The input integers range of 0 to 1 is the logistic function, sometimes referred to as the sigmoid function. Numbers around 0 suggest a low probability, whereas numbers near one indicate a high chance.

$$F(z) = \frac{1}{1+e^{-z}} \tag{9}$$

where, $F(z)$ denotes sigmoid function and e represent Euler number.

Integrated Gradients

IGs are a well-liked attribution-based XAI method for assessing if words significantly affect the target prediction (Moraliyage et al. 2025). In attribution-based techniques, DNN predictions are used to determine the relevance of the input qualities that explain the model predictions. Any differentiable DNN model can be used with IGs, a local explainability method that offers several advantages over alternative attribution methods. The IG XAI technique is also easier to implement in DNNs than earlier approaches. Establishing a reference baseline is necessary to obtain feature attributions using IGs. Compared to this baseline, IGs provide feature attributions that lack useful content. The input features are defined in Equation (10)

$$IG_j(s, s') ::= (s_j - s'_j) \times \int_{\alpha=0}^{1} \frac{\partial p(s' + \alpha \times (s - s'))}{\partial s_j} d\alpha \tag{10}$$

Define that $p : R^n \to [0,1]$ to denote the deep neural network and let $s \in R^n$ as the input while $s' \in R^n$ is the input of baseline. IG is calculated by summing all gradients assessed along a straight line from the input s to the baseline s'. IG is the route integral of the gradients calculated from the input s to the baseline s' along a straight-line path. Equation (10) describes the IG for an input s and baseline s' along the j-th dimension. The term $\frac{\partial p(s)}{\partial s_j}$ denotes the model p gradient with regard to the j-th dimension. This reference baseline attempts to produce a high entropy prediction with increased uncertainty. By adding the input features to the reference baseline, which interpolates towards the entire input, the uncertainty resulting from the absence of the input features is thus reduced. IG is then computed by summing these gradients.

3.6 LwCIG optimized using GVROA

LwCIG is optimized using a hybrid GVOA-ROA structure for early diabetes prediction. To integrate the global search capability of the GVOA with the fine-tuning and precision of the ROA, the efficient exploration of the feature space, and the demonstrated convergence toward optimal predictive parameters are demonstrated. The hybrid approach improves the performance model by balancing exploration and exploitation, being rapid, and enabling accurate and scalable detection of early diabetic pattern(s) using both clinical and lifestyle data.

GVOA

This paper discusses GVOA (Hasan et al., 2025) as a new and promising approach to optimizing processes. It is based on the intelligent foraging behaviour of griffon vultures and uses a mathematical formulation in its algorithm to show how it successfully solves complex optimization problems.

Phase 1: Initialization

This step starts with the initial population in GVOA: the population consists of a random set of possible solutions to the optimization problem, with each solution being an entry in the form of a location of a griffon vulture in the

search space, as indicated in (11). There, $\dim$ is the dimensionality of the problem, $HW_{pop.\dim}$ denotes the number of potential solutions, and pop denotes the number of candidate solutions. The term $\dim$ denotes the pop^{th} griffon vulture, which is located in the $\dim^{th}$ dimension. This strategy guarantees that the griffon vultures' locations are evenly dispersed around the search space.

$$K = \begin{bmatrix} HW_{1.1} & HW_{1.2} & \cdots & \cdots & HW_{1.\dim} \\ HW_{2.1} & HW_{2.2} & \cdots & \cdots & HW_{2.\dim} \\ \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ HW_{pop.1} & HW_{pop.w} & \cdots & \cdots & HW_{pop.\dim} \end{bmatrix} \tag{11}$$

Equation (12) states that the starting positions of the GVOA population within the search space are randomly generated at the beginning of the process. In this case, $Rand$ represents a random number between 0 and 1, while vc and mc represent the problem's upper and lower bounds.

$$K = mc + Rand_1 \times (vc - mc) \tag{12}$$

Phase 2: Fitness

The optimal solution is calculated with the fitness function can be represent as Equation (13)

$$f(y) = \sum_{j=1}^{o} \omega_j . o_j(y) \tag{13}$$

where, $o_j(y)$ denotes the objective terms of the accuracy, ω_j represent weights.

Phase 3: Following behavior

Griffon vultures display a unique "following behavior" in which individuals with less information follow those with more to locate food. Some vultures are uninformed, while others possess varying degrees of information based on past experiences. Less knowledgeable people follow those with greater objective knowledge, while ignorant vultures randomly follow knowledgeable ones. Using its own information, the most knowledgeable vulture takes the lead and advances straight toward the carcasses. This behavior enables exploration of the search space in a variety of effective ways, thereby improving the GVOA's global search capabilities. During the exploration phase, which is dominated by exploration, uneducated griffon vultures follow knowledgeable individuals, dispersing across the search area rather than converging on a single location. On the other hand, it signifies exploitation when the most knowledgeable vulture approaches the carcass directly. Equation (14), divides the Griffin vultures (agents) into two groups based on the information given: informed and uninformed. The most knowledgeable agent is the one with the lowest objective function value among the informed agents. Furthermore, $carcass$ symbolizes the finest answer discovered by vultures worldwide thus far.

$$\inf ormedGV, uninf ormedGV = sort(fit, ascend)[1:M/2], [M/2+1:M] \tag{14}$$

The status of each Griffon Vulture as either informed or misinformed is then calculated using Equation (15) $Y_j(u) \in uninf ormedGV$. The random informed griffon vulture $SelGV$ is chosen selecting the first vulture that is uninformed. If the vulture is knowledgeable, the second instance of Equation (12) is used to select, if available, a more knowledgeable Griffin vulture.

$$SelGV = \begin{cases} rand\left(\left\{Y_c \middle| c \in \inf ormedGV\right\}\right), if Y_j(u) \in uninf ormedGV \\ rand\left(\left\{Y_c \middle| fit_c < fit_{j,} c\right\}\right), j, c \in \inf ormedGV, else \end{cases} \tag{15}$$

The variable is introduced in Equation (15) in order to preserve equilibrium between exploration and exploitation. A random number between 0 and 2 is assigned when the temperature is less than $\frac{MaxIt}{3}$, which encourages aging

and a greater emphasis on exploitation. The focus is placed more on exploration when c is given a random value between 0 and 4 when the temperature is greater than or equal to $\frac{MaxIt}{3}$.

$$c = 2 \times rand + \left(it \geq \frac{MaxIt}{3} \right) \times \left(2 \times rand \right) \tag{16}$$

In this case, $MaxIt$ denotes the maximum number of permitted iterations. The technique uses Equation (17) to establish a new location for the vulture. This new location replaces the prior one, which is captured in Equation (18), if it increases the objective function value.

$$Y_{GC}(o+1) = Y_j(o) + c.\left(carcassPos - sO \times Y_j(o)\right) \tag{17}$$

Assume $Y_{GC}(o+1) = Y_j(o+1)$ and substitute in Equation (17)

$$Y_j(o+1) = Y_j(o) + c.\left(carcassPos - sO \times Y_j(o)\right) \tag{18}$$

$$Y_j(o+1) = Y_j(o) + c.\ carcassPos - sO \times Y_j(o) \tag{19}$$

$$Y_j(o+1) = Y_j(o)[1+c] + c.carcassPos - csO \tag{20}$$

$$Y_j(o)[1+c] = Y_j(o+1) - c.\ carcassPos - csO \tag{21}$$

$$Y_j(o) = \frac{1}{1+c}\left[Y_j(o+1) - c.\ carcassPos - csO\right] \tag{22}$$

The newly determined location for the current Griffin Vulture based on the first phase following behavior of the GVOA algorithm, is represented by Y_{GC}.

Integration of Revolution Optimization Algorithm into GVROA

Ideology is the fundamental set of ideas that guides the goals and strategies of revolutionary movements follow during revolutions. It serves as a framework that encourages and motivates people to take part in transformational activities by providing hope and a vision for a better future. The success of the movement largely depends on how well the populace receives the leader's philosophy. A leader may foster unity, inspire followers, and ignite a revolutionary spirit among the populace by effectively advancing a widely recognized and motivating philosophy.

In this stage of the ROA (Hamadneh et al. 2025), the principle of growing self-awareness is used to update the population members. This stage simulates how people progressively gain knowledge from their experiences and adjust their behavior to produce better outcomes over time. This stage ensures that the algorithm investigates regions with greater potential, optimizing the exploitation of prospective solutions, by focusing on the surroundings of known solutions.

$$Y_j(o+1) = Y_j(o) + s\left(Y_j^{old} - Y_j(o)\right) \tag{23}$$

$$Y_j(o+1) = Y_j(o)[1-s] + sY_j^{old} \tag{24}$$

Substitute (22) in (24)

$$Y_j(o+1) = \frac{1}{1+c}\left[y_j(o+1) - c.carcassPos - csO\right][1-s] + sY_j^{old} \tag{25}$$

$$Y_j(o+1) - \frac{Y_j(o+1)[1-s]}{1+c} = \frac{1}{1+c}\left[-c\,carcassPos - csO\right][1-r] + sY_j^{old} \tag{26}$$

$$Y_j(o+1)\left[\frac{1+c-1+s}{1+c}\right] = \frac{1}{1+c}\left[-c\,carcassPos - csO\right][1-r] + sY_j^{old} \tag{27}$$

$$Y_j(o+1)\left[\frac{c+s}{1+c}\right] = \frac{1}{1+c}\left[-c\,carcassPos - csO\right][1-r] + sY_j^{old} \tag{28}$$

$$Y_j(o+1) = \frac{1+c}{c+s}\left[\frac{1}{1+c}\left[-c\,carcassPos - csO\right][1-s] + sY_j^{old}\right] \tag{29}$$

Equation (24) denotes the j^{th} members newly determined position of ROA. The j^{th} dimension of the members' position in the previous iteration is represented by the symbols Y_j^{old} . This framework, which enhances early diabetic prediction of GVOA-ROA, integrates with GVROA. To ensure interpretability, IG are employed, which highlight the contribution of individual input features models predictions. Using the GVROA method improves how quickly we can optimize the process of choosing the best option, helping us identify and implement greater options while also providing the decision-making process with better trustworthiness and transparency for use in clinical applications. In addition, this model will achieve high accuracy and provide a high level of explicitness associated with that accuracy.

Phase 4: Group Foraging

Griffon vultures (Gyps fulvus) forage for food in a complex manner by using their collective knowledge to find food sources. When they see a dead animal, instead of immediately flying to it, they usually wait until many other vultures have also seen it;, to avoid the risk of dying (an energy-saving behaviour). The leadership within this particular behaviour is dynamic; an informed vulture leads others to the dead animal while maintaining its own leadership until new information becomes known to the rest of the vultures. Each vulture moves toward the dead animal based on random numbers. If the randomly generated number is less than or equal to 0.2, then the vulture moves in the same direction as other vultures; otherwise it moves toward the dead animal. This process illustrates how to explore the environment in accordance with social cues and how to use the information from the other vultures.

Phase 5: Vulture independent search

Alongside foraging cooperatively, vultures explore alone by using their excellent eyesight, looking for food within a confined, predetermined area. The model simulates this behavior through a linear decay parameter that reduces the randomness of their movement over time, facilitating a more directed and refined search. The vultures each search around their location, adjusting their movements to indicate they are seeking a better carcass position if they find one. This phase provides a balanced combination of exploration and exploitation, improving the efficiency of the local search and allowing the model to converge to optimal solutions.

Phase 6: Near carcass scout

Before feeding on the carcass, vultures conduct a scouting flight to assess a safe approach and minimize predation risk. The parameter is included in the model to provide early variation for exploratory movement and gradual stabilization as the iterations progress. The vultures' search is centered on the carcass location, and the addition of a Gaussian perturbation simulates more realistic, stochastic adjustments to their movement. If a new candidate location has a better objective value, the model will also update the carcass location. This phase also provides a balanced combination of exploration and exploitation, enabling more efficient overall search.

Phase 7: Re-evaluate the fitness

The original fitness function has been recalculated with the purpose of minimizing error. Each aggregate objective term is weighted by its importance.

Phase 8: Termination

Upon completion of achieving the best possible solution (from which to work), all processes will be stopped. Table 2 shows a Pseudocode for GVROA.

Table 2: Pseudocode for GVROA

START

Step1. Initialization:

For each vulture $i = 1$ to pop :

For each dimension $j = 1$ to dim :

$K[i][j] = mc + Rand() * (vc - mc)$

Evaluate fitness $f(K[i])$ for each vulture

$carcassPos$ = best position found

$carcassFitness$ = corresponding best fitness

Step 2: Exploration

Main Loop (iteration = 1 to MaxIt) :

Sort vultures based on fitness

Divide into two groups: informedGV, uninformedGV

For each vulture j :

Select informed vulture SelGV

If iteration < MaxIt / 3 :

$c = \text{Rand}() * 2$

else:

$c = \text{Rand}() * 4$

$GCY = \text{carcassPos} + c * (\text{carcassPos} - \text{SelGV})$

If $f(GCY) < f(K[j])$

$K[j] = GCY$

Step 3: Revolution-based Exploitation Phase:

For each vulture j :

$Y\_old = K[j]$

$s = \text{Rand}()$

$K[j] = Y\_old + s * (Y\_old - Y[j])$

$K[j] = Y\_old - s * (Y\_old - Y[j])$

Step 4: Recalculate fitness

For each vulture j :

Evaluate $f(K[j])$

If $f(K[j]) < \text{carcassFitness}$:

$\text{carcassFitness} = f(K[j])$

$\text{carcassPos} = K[j]$

Step 5: Termination

Return carcassPos and carcassFitness

STOP

4. Experimental results

Results demonstrate that the proposed early diabetes risk prediction model achieves high accuracy, precision, recall, and F1-score, outperforming existing methods. Hybrid GVOA-ROA integrates optimization, enhanced convergence speed, and model generalization. IG highlights key risk factors that contribute to the predictions. The model improves effectiveness and reliability for early diabetic prediction.

4.1 System and Algorithm Configuration

The System ran Windows 10 with a 64-bit kernel and it was powered by a processor with 3GB of memory and 32GB of RAM. Python version 3.12.7 was used in a jupyter Notebook development environment. The training model consisted of 120 iterations and used 8GB of storage. Essential software dependencies included various python libraries required for development and execution. This configuration will provide a solid foundation for testing, and there will be significant resource efficiency to implement the model.

4.2 Distribution of health conditions among adults

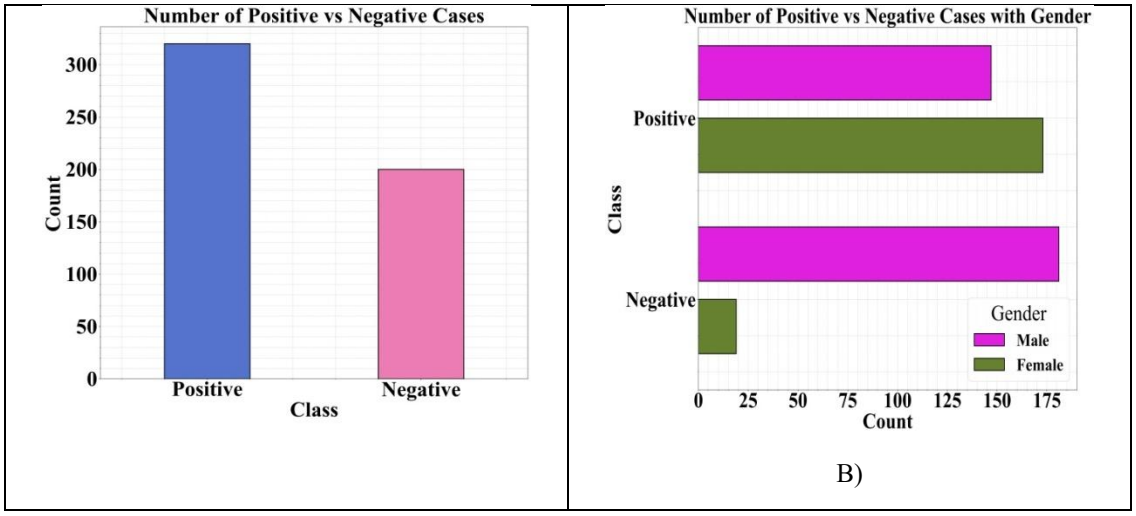

Figure 3: Distribution of positive and negative health condition among adults A) Class distribution B) Gender-class distribution

The total number of individuals classified as being positive or negative based on a particular characteristic for the conditions illustrated in Figure 3. As shown in Figure 3A), the number of individuals with a positive characteristic is approximately 320 (positive), while 200 individuals have a negative characteristic. This means there are more than twice as many persons in the data with a positive characteristic as with a negative characteristic. The unequal distribution of classes may affect the predictive model, creating a need for class balancing (e.g., oversampling and/or class weighting). Figure 3B) expands upon this classification group in further detail with respect to gender. It reveals that among individuals with a positive characteristic, the vast majority are male, whereas among negative individuals, the vast majority is female. Male individuals account for the majority of positive cases, while females account for the majority of negative cases. Classifying participants by gender is an important line of inquiry regarding trends and risks associated with having this condition. Thus, it may justify additional gender-targeted analyses or considerations when developing predictive models or delivering health services to affected individuals.

4.3 Histogram of age distribution

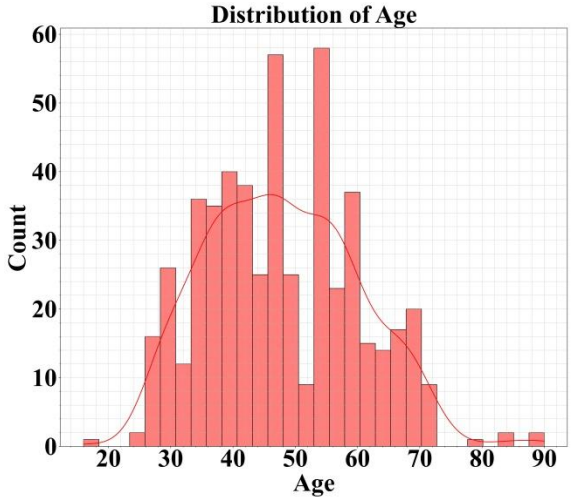

Figure 4: Distribution of age

The distribution of individuals by age in Figure 4 shows how many individuals are present in each age group. Most of the ages in this sample are between 40 and 60; the largest number of individuals is at age 50, indicating that the sample population is composed primarily of middle-aged individuals. The distribution shows a slight right skew and a smaller number of individuals that fall in the older age brackets, over age 70. There is also a gradual increase in the frequency of individuals from 30 to 50; however, there is a resulting decrease in the frequency of individuals from age 50 to 70. The shape of the curve overlaid on the age distribution is typical of a bell-shaped distribution, with some variability. Therefore, the results approximate a normal distribution with the central tendency lying within the middle age group.

4.4 Distribution of Health Features in Diabetic and Non-Diabetic Individuals

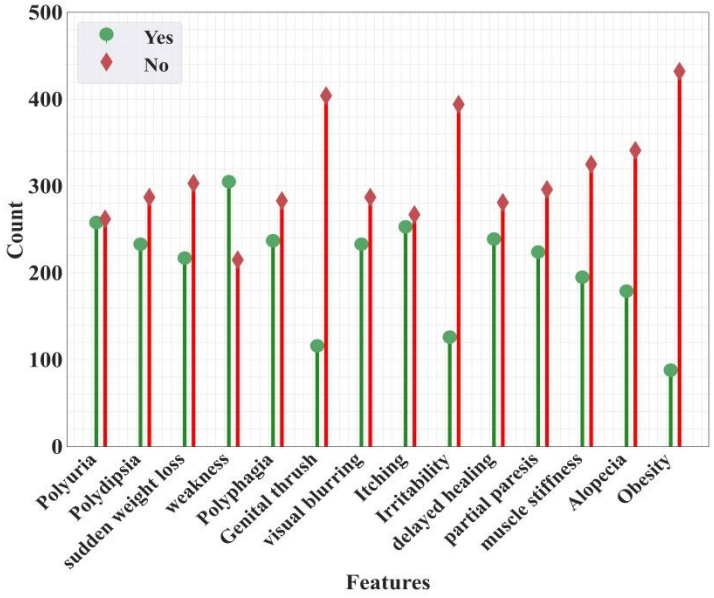

Figure 5: Features of diabetics and non-diabetics

Figure 5 shows the comparison of various features between diabetic mention “yes” and non-diabetic mention “no” individuals. High counts for diabetics are observed in polyuria, polydipsia, sudden weight loss, weakness, visual blurring, itching, delayed healing, and obesity, each with counts above 300, some nearing or exceeding 400. These features are strongly associated with diabetes. In contrast, features like genital thrush, alopecia, and polyphagia have relatively moderate counts around 250 to 300 for diabetics. Low counts for diabetics below 200 are observed for muscle stiffness and irritability, whereas the non-diabetic group shows a similar or even slightly higher prevalence.

Obesity and visual blurring show the most significant difference between the two groups. This indicates that certain symptoms are strong indicators of diabetes, while others show overlapping trends.

4.5 Confusion matrix of features in diabetes dataset

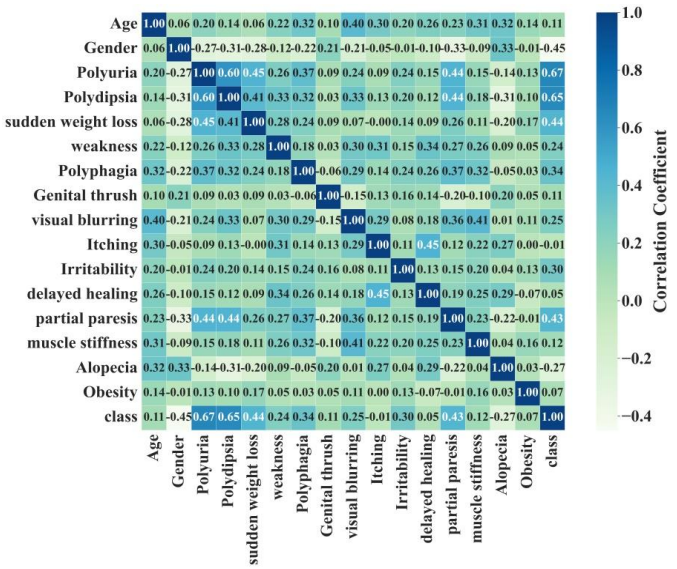

Figure 6: Confusion matrix of diabetes dataset

Figure 6 shows a correlation matrix for the early diabetes dataset, the relationships between various clinical features and the diabetic condition. The correlation coefficient matrix depicts the degree of correlation between two independent variables; values can range from -0.4 to +1.0. A positive correlation coefficient close to +1 indicates a strong positive correlation, and any negative correlation coefficient indicates an inverse correlation. Polyuria, polydipsia, and sudden weight loss display a strong positive correlation with the diabetes class label (indicating the importance of these features for diagnostic purposes). In contrast, variables such as gender and alopecia have very weak correlations. The correlation matrix will determine which features correlates strongly with early diabetes detection.

4.6 Optimization Comparison of fitness function

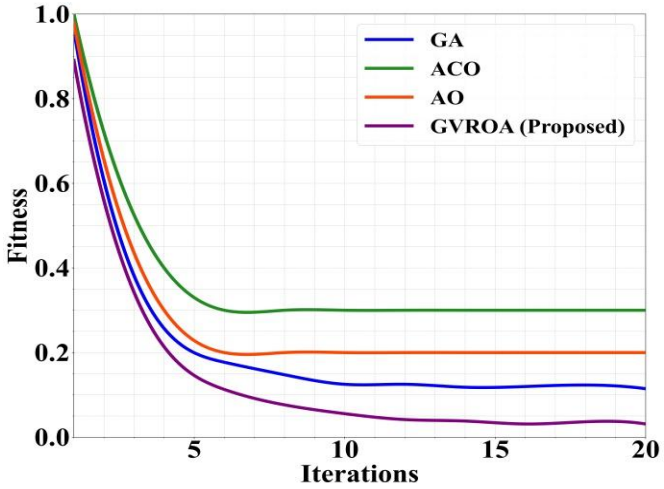

Figure 7: Fitness comparison of existing method with proposed method

Figure 7 compares the fitness performance of Genetic Algorithm (GA) (Bülbül 2024), Ant Colony Optimization (ACO) (Jain et al. 2024), Adam Optimization (AO) (Daliya and Ramesh 2025), and the proposed GVROA method over 20 iterations. All algorithms show a decrease in fitness, indicating convergence towards optimal solutions. ACO

converges slowly and stabilizes at a higher fitness level, reflecting less effective optimization. GA and AO perform better but still fall short in achieving lower fitness values. The proposed GVROA method consistently outperforms others by rapidly decreasing fitness within the first iteration. It reaches the lowest fitness value and maintains stability throughout. This highlights its faster convergence speed and improved efficiency. GVROA's robust performance demonstrates its effectiveness in optimizing complex problems. Furthermore, the proposed method achieves superior results compared to existing techniques.

4.7 Performance evaluation with comparison metrics

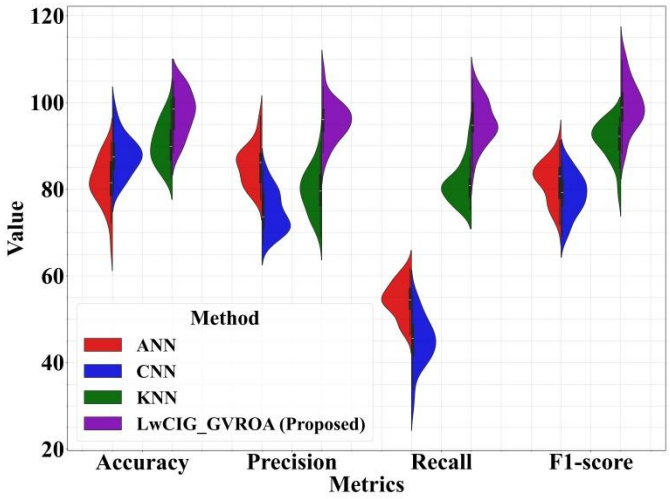

Figure 8: Comparison of existing methods with proposed methods

In Figure 8, the comparison of precision, recall, and F1 score is shown between the current ANN (Rastogi & Bansal 2023), CNN (Zhou et al. 2023), KNN (Alnowaiser 2024), and the suggested LwCIG_GVROA model. The LwCIG_GVROA model performs significantly better than any of the other models with an accuracy of (95.20%), precision of (96.22%), recall of (97.12%), and F1 score of (99.65%). ANN has an accuracy of (79.49%), a precision of (85%), a recall of (56%), and an F1 score of (83%). CNN has an accuracy of (89.28%), a precision of (75%), a recall of (44%), and an F1 score of (78%). KNN has an accuracy of (91.17%), a precision of (80%), recall of (82%), and an F1 score of (92%) demonstrating the effectiveness of the hybrid optimization and integrated lightweight convolutional model. In addition to improved recall and F1 score, the results suggest more true positives and fewer false negatives. Additional adjustments could include developing interpretable modules to better understand the model better, using larger datasets in real-world settings, using domain adaptation methods to favour culturally diverse populations, and developing federated learning capabilities across sites to improve the model's generalizability and clinical utility. Reducing the computational load on memory and processors will ensure this can be used in real time in mobile health applications.

4.8 *Ablation Study*

Table 3: Ablation study

Methods	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
VGG-16	93	93.5	87.23	90.6
BiLSTM	91.25	92.33	90.55	94.43
LSTM	94.10	95	93.6	97.31
LwCIG	**95.20**	**96.22**	**97.12**	**99.65**

In Table 3, VGG-16, BiLSTM, LSTM, & our proposed LwCIG model are evaluated against each other on several critical performance metrics. The results indicate that our proposed model LwCIG has significantly higher performance across all metrics than the other three models evaluated herein. The VGG-16 & BiLSTM have very similar precision & recall but relatively lower F1 scores and overall accuracy compared to the other two models evaluated. The LSTM outperforms the VGG-16 & BiLSTM, achieving 95%, recall, 93.6%, and an F1-score of almost 97.31%. On the other hand, the proposed LwCIG model outperforms all other models on each of the evaluated performance metrics with the highest overall accuracy (95.20%), the highest precision (96.22%), recall (97.12%), & with the highest F1-score of 99.65% resulting from greater overall balance in the precision & recall metrics & offering the best performance as a classification model.

4.9 Discussion

The LwCIG_GVROA model is designed to predict early diabetes risk by robustly preprocessing data, carefully selecting features, augmenting data with CTGAN, and using interpretable deep learning. This combination will provide accurate predictions, high explainability, and low computational cost. In a real-time clinical setting, the model will provide proactive screening for patients at risk of diabetes using available clinical and symptomatic data enabling early intervention. The lightweight design will enable deployment on low resource systems, such as portable diagnostic tools used in rural clinics. Additionally, the use of synthetic data generated with CTGAN will address class imbalance and improve the model's across diverse patient populations. The combination of MI and Chi-Squared feature selection provides the model with the most relevant features for prediction while reducing noise and improving overall accuracy of the diagnosis. The use of information gain will clinicians obtain transparent explanations of decision-making factors. Further, the use of GVROA for model optimization will help the model adapt to the complex nature of real-world health care data. Overall, the framework will be scalable and capable of providing clinical decision support for diabetes management.

4.9.1 Interdisciplinary Implications: Behavioral, Policy, and Organization perspectives

The IG uses several methods, including behavioral-level risk factors, engaging parents in preventive techniques, establishing trust between providers and patients, and increasing awareness of diabetes risk, at the foundation of each method. By identifying polydipsia, polyuria, and obesity, clinicians will have access to helping counsel individuals on how to change their lifestyles over time. From a policy perspective, using early-stage predictors to identify diabetes risk enables the development of preventive health programs for systematic risk stratification and large-scale screening across the general population. These tools can significantly decrease the burden of disease and long-term treatment costs, as well as increase the health equity for those with limited access to quality medical care. From an organizational perspective, the lightweight architecture enables primary care professionals and their associated records to share data through information systems. This capability reduces the diagnostic responsibilities placed on all professionals providing care to their patients while coordinating across all professionals supporting the delivery of care. Lastly, the model's computational efficiency enables deployment at rural health clinics and community health centers, strengthening health professionals' ability to offer preventive health services.

4.9.2 Data Governance, Ethical Considerations, and Real-World Implementation

In AI-driven diabetes risk assessment systems, regulatory compliance, data privacy, and data security are critical factors in the healthcare domain. Clinical datasets contain PHI and must also conform to the same standards used to protect personal health information (PHI), including access control, anonymization, and secure data storage. Additionally, ethical considerations apply to the use of predictive models. At the same time, the Integrated Guidance (IG) assists clinicians in interpreting and verifying the outputs of the predictive models they will use to inform their decision-making. The IG encourages clinicians to use their clinical judgment to reduce the risk of dependence on predictive models, and to consider the potential biases associated with the use of predictive modeling tool when imbalanced demographic data is used. Furthermore, the lightweight architecture of the predictive modeling tool we propose supports practical implementation within clinical decision support systems and electronic health records (EHRs). The Human-in-the-loop will help promote accountability and patient safety, as the clinician will have the final decision-making authority over the use of predictive modeling in diabetes risk assessment. To maintain trust in predictive models, outcomes will require continuous model evaluation and monitoring of the clinical effectiveness of predictive models after implementation.

5. Conclusion

The hybrid of GVROA and LwCIG can create an effective, easily interpretable model to predict risk for early-onset diabetes. The metrics achieved with this model include high performance, effective data preprocessing and cleaning, relevant features selected with CTGAN to augment the minority class, and a lightweight CNN produces results quickly for real-time use in the medical field. Additionally, the use of IG in the model provides greater transparency into how prediction results are derived. Compared to other optimization methods, the new GVROA method consistently reduced fitness values to 0.1 after only one iteration. This fast convergence enabled accelerated training while maintaining the model's accuracy and stability. The proposed model demonstrated robustness, scalability, and reliability when managing complex clinical data from numerous sources. The model supports river methods and clinical workflows for community-based diabetes prevention initiatives, where predictive accuracy must balance the feasibility of deployment and interpretability. The model provides a clinically relevant framework to support accurate, explainable medical decision making; however, it is dependent upon the quality of the imputation process as well as the quality of the synthetic data that is generated thus potentially impeding performance when the underlying dataset is highly noisy/incomplete, The model's architecture may also be an impediment in that it may require retraining for different diseases. In addition to providing a means for predicting future individual behaviours relating to their health, this method also supports behaviour modification, assists in developing preventive health policy and aids in successfully implementing organizational change within health care systems. The method, therefore, creates opportunities to develop scalable and sustainable diabetes prevention programmes by bridging AI with the policy and the clinical behaviour realms. The responsible deployment of AI-based risk prediction systems requires ethical transparency, robust data governance, and compliance with regulations. In addition, this method supports scalability and has been demonstrated to be trustworthy in actual health care settings by integrating clinician oversight, implementation feasibility, and explainability.

The future of this study will extend to a multi-disease risk model using multi-label classification. CTGAN could be enhanced with privacy-preserving methods such as differential privacy to improve data security. Using temporal health record data could provide additional information to improve the model's predictive performance. Techniques for compressing models can be explored and applied to reduce resource use when deploying the model on embedded devices. Integration with wearable IoT devices will be another area of promise. Finally, large-scale clinical trials should be conducted to verify the model and make it more widely available in the healthcare system.

Acknowledgements

None

Disclosure of interest

No potential competing interest was reported by the authors

Funding

No funding was received

Compliance with Ethical Standards

1. **Potential Conflict of Interest Disclosure:**

   None of the authors have any potential conflicts of interest.

2. **Declaration on Human and Animal Rights**

   a. Ethical Approval: All relevant national and/or academic regulation for the use and care of animals was strictly adhered to.

3. **Informed Consent**

   Formal consent is not essential for this particular type of research.

Replication of results

No results are presented

Data availability statements

As no datasets were created or examined for this research, data sharing is not appropriate.

Authors' contribution

Every author has made an equal contribution to the work. All authors reviewed the manuscript.

References

1. Ahakonye LAC, Nwakanma CI, Lee JM, Kim DS. 2023. SCADA intrusion detection scheme exploiting the fusion of modified decision tree and chi-square feature selection. Internet Things. 21:100676.
2. Ahmed BM, Ali ME, Masud MM, Naznin M. 2024. Recent trends and techniques of blood glucose level prediction for diabetes control. Smart Health. 32:100457.
3. Alnowaiser K. 2024. Improving healthcare prediction of diabetic patients using KNN imputed features and tri-ensemble model. IEEE Access. 12:16783–16793.
4. Bülbül MA. 2024. A novel hybrid deep learning model for early stage diabetes risk prediction. J Supercomput. 80(13):19462–19484.
5. Daliya VK, Ramesh TK. 2025. A cloud based optimized ensemble model for risk prediction of diabetic progression—An Azure machine learning perspective. IEEE Access.
6. Doğru A, Buyrukoğlu S, Arı M. 2023. A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. Med Biol Eng Comput. 61(3):785–797.
7. Early Stage Diabetes Risk Prediction Dataset. 2025. Retrieved from https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset
8. El-Sofany H, El-Seoud SA, Karam OH, Abd El-Latif YM, Taj-Eddin IA. 2024. A proposed technique using machine learning for the prediction of diabetes disease through a mobile app. Int J Intell Syst. 2024(1):6688934.
9. Gowthami S, Reddy RV, Ahmed MR. 2024. Exploring the effectiveness of machine learning algorithms for early detection of Type-2 diabetes mellitus. Meas Sensors. 31:100983.
10. Gündoğdu S. 2023. Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique. Multimed Tools Appl. 82(22):34163–34181.
11. Haldorai A, Murugan S, Balakrishnan M. 2024. Risk prediction of maternal health by model analysis using artificial intelligence. In: Artificial Intelligence for Sustainable Development. Cham: Springer Nature Switzerland:125–138.
12. Hamadneh T, Batiha B, Gharib GM, Montazeri Z, Dehghani M, Aribowo W, Noori HM, Jawad RK, Ibraheem IK. 2025. Revolution optimization algorithm: A new human-based metaheuristic algorithm for solving optimization problems. Int J Intell Eng Syst. 18(2).
13. Hasan DO, Mohammed HM, Abdul ZK. 2025. Griffon vultures optimization algorithm for solving optimization problems. Expert Syst Appl. 276:127206.
14. Hossain MJ, Al-Mamun M, Islam MR. 2024. Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. Health Sci Rep. 7(3):e2004.
15. Jadon AS, Kaushik MP, Anitha K, Bhatt S, Bhadauriya P, Sharma M. 2024. Types of diabetes mellitus, mechanism of insulin resistance and associated complications. In: Biochemical Immunology of Diabetes and Associated Complications. Academic Press:1–18.
16. Jain H, Chourey A, Chaure R, Shrivastava R. 2024. An innovative approach for enhanced pattern extraction utilizing ant colony optimization. In: Artificial Intelligence and Information Technologies. CRC Press: 343–347.
17. Khalifa M, Albadawy M. 2024. Artificial intelligence for diabetes: Enhancing prevention, diagnosis, and effective management. Comput Methods Programs Biomed Update. 5:100141.
18. Khosravi H, Farhadpour S, Grandhi M, Raihan AS, Das S, Ahmed I. 2024. Strategic data augmentation with CTGAN for smart manufacturing: Enhancing ML predictions of paper breaks in pulp-and-paper production. Manuf Lett. 41:1312–1323.
19. Khunti K, Zaccardi F, Amod A, Aroda VR, Aschner P, Colagiuri S, Mohan V, Chan JC. 2025. Glycaemic control is still central in the hierarchy of priorities in Type 2 diabetes management. Diabetologia. 68(1):17–28.
20. Liu R, Qu Z, Feng Y, Bai L, Liu X, Fan X, Liu X, Zhao L. 2025. Progress in the treatment of vascular complications in Type 2 diabetes by finerenone in combination with RAS inhibitors/SGLT-2i. J Diabetes Its Complications. 108981.
21. Moraliyage H, Kulawardana G, De Silva D, Issadeen Z, Manic M, Katsura S. 2025. Explainable artificial intelligence with integrated gradients for the detection of adversarial attacks on text classifiers. Appl Syst Innov. 8(1):17.
22. Ojurongbe TA, Afolabi HA, Oyekale A, Bashiru KA, Ayelagbe O, Ojurongbe O, Abbasi SA, Adegoke NA. 2024. Predictive model for early detection of Type 2 diabetes using patients' clinical symptoms, demographic features, and knowledge of diabetes. Health Sci Rep. 7(1):e1834.
23. Olabanjo O, Wusu A, Olabanjo O, Asokere M, Afisi O, Akinnuwesi B. 2025. A novel deep learning model for early diabetes risk prediction using attention-enhanced deep belief networks with highly imbalanced data. Int J Inf Technol. 1–23.
24. Olisah CC, Smith L, Smith M. 2022. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Comput Methods Programs Biomed. 220:106773.
25. Oliullah K, Rasel MH, Islam MM, Islam MR, Wadud MAH, Whaiduzzaman M. 2024. A stacked ensemble machine learning approach for the prediction of diabetes. J Diabetes Metab Disord. 23(1):603–617.
26. Pokhrel A, Chong KT, Tayara H. 2025. Therapeutic potential of curcuminoids in Type 2 diabetes mellitus (T2DM): Insights from network pharmacology, molecular docking, and dynamics simulations. Food Biosci. 106406.
27. Rastogi R, Bansal M. 2023. Diabetes prediction model using data mining techniques. Meas Sensors. 25:100605.

28. Sheth KA, Upreti C, Prusty MR, Satapathy SK, Mishra S, Cho SB. 2024. Time-frequency transformation integrated with a lightweight convolutional neural network for detection of myocardial infarction. BMC Med Imaging. 24(1):1–15.
29. Stoleru OA, Burlec AF, Mircea C, Felea MG, Macovei I, Hancianu M, Corciovă A. 2024. Multiple nanotechnological approaches using natural compounds for diabetes management. J Diabetes Metab Disord. 23(1):267–287.
30. Su SS, Li LY, Wang Y, Li YZ. 2023. Stroke risk prediction by color Doppler ultrasound of carotid artery-based deep learning using Inception V3 and VGG-16. Front Neurol. 14:1111906.
31. Sun Q, Cheng X, Han K, Sun Y, Ren H, Li P. 2025. Machine learning-based assessment of diabetes risk. Appl Intell. 55(2):1–13.
32. Teixeira PF, Battelino T, Carlsson A, Gudbjörnsdottir S, Hannelius U, von Herrath M, Knip M, Korsgren O, Elding Larsson H, Lindqvist A, Ludvigsson J. 2024. Assisting the implementation of screening for Type 1 diabetes by using artificial intelligence on publicly available data. Diabetologia. 67(6):985–994.
33. Xu S, Dai Y, Yan C, Sun Z, Huang R, Guo D, Yang G. 2025. On the preprocessing of physics-informed neural networks: How to better utilize data in fluid mechanics. J Comput Phys. 113837.
34. Yadalam PK, Arumuganainar D, Ronsivalle V, Di Blasio M, Badnjevic A, Marrapodi MM, Cervino G, Minervini G. 2024. Prediction of interactomic hub genes in PBMC cells in Type 2 diabetes mellitus, dyslipidemia, and periodontitis. BMC Oral Health. 24(1):385.
35. Yu F, Guan J, Wu H, Wang H, Ma B. 2025. Multi-population differential evolution approach for feature selection with mutual information ranking. Expert Syst Appl. 260:125404.
36. Zeinalnezhad M, Shishehchi S. 2024. An integrated data mining algorithms and meta-heuristic technique to predict the readmission risk of diabetic patients. Healthcare Anal. 5:100292.
37. Zhou H, Xin Y, Li S. 2023. A diabetes prediction model based on Boruta feature selection and ensemble learning. BMC Bioinformatics. 24(1):224.