KMANPUB | Health Nexus Vol. 5 No. 1 (2027): 1-24

Article history:
Received 21 May 2026
Revised 24 August 2026
Accepted 31 August 2026
Initial Publication 09 September 2026
Final Publication 01 January 2027

# A Diabetes Diagnosis Approach by Feature Selection Using Group Learning Optimization (GLO) Algorithm and Apache Spark Distributed Processing Technology

Nasrin Evazzadeh Mohammadian[1], Mehdi Jafari Shahbazzadeh[2*], Amir Sabbagh Molahosseini[1]

[1] Department of Computer Engineering, Ke.C., Islamic Azad University, Kerman, Iran
[2] Department of Electronic and Electrical Engineering, Ke.C., Islamic Azad University, Kerman, Iran

*** Corresponding author email address**: mjafari@iauk.ac.ir

A r t i c l e  I n f o

**Article type:**
*Original Research*

**How to cite this article:**
Evazzadeh Mohammadian, N., Jafari Shahbazzadeh, M., & Molahosseini, A. S. (2027). A Diabetes Diagnosis Approach by Feature Selection Using Group Learning Optimization (GLO) Algorithm and Apache Spark Distributed Processing Technology. *Health Nexus*, *5*(1), 1-24.
https://doi.org/10.61838/kman.hn.6008



A B S T R A C T

This study aimed to develop and evaluate an accurate and computationally efficient diabetes diagnosis framework integrating GAN-based class balancing, binary Group Learning Optimization (GLO) feature selection, and Apache Spark distributed machine-learning classification. The study used the Pima Indians Diabetes Dataset containing 768 records with eight diagnostic features and a binary diabetes outcome. Data were normalized to the [0,1] range, and a Generative Adversarial Network was used to balance minority and majority classes. A binary version of the GLO algorithm was developed for feature selection by minimizing a fitness function combining classification error and the number of selected features. The dataset was divided into 70% training and 30% testing subsets. A multilayer perceptron was used to evaluate candidate feature vectors during optimization. The reduced-dimensional data were subsequently processed within an Apache Spark master-worker architecture and classified using Decision Tree, Random Forest, Support Vector Machine, and majority voting. GLO was additionally evaluated against TLBO, WOA, SHO, WSA, and HHO using benchmark optimization functions. GLO achieved the best mean optimization rank of 1.86 and the lowest local-optimum entrapment rate of 7.5%. In feature-selection experiments, GLO achieved 89.76% accuracy, 88.60% sensitivity, and 88.41% precision, outperforming CSA, GWO, GOA, PSO, and GA in diagnostic accuracy. GAN balancing improved Decision Tree, Random Forest, and SVM accuracy to 93.21%, 94.62%, and 91.63%, respectively, while majority voting achieved 93.92% accuracy, 93.67% sensitivity, and 92.84% precision. In the distributed Spark implementation, Random Forest achieved the highest diagnostic performance, with 98.97% accuracy, 98.64% sensitivity, and 98.62% precision. Computational acceleration increased with the number of Spark clusters, with Decision Tree attaining a maximum speedup of 12.56. Integrating GAN-based balancing, binary GLO feature selection, and Apache Spark distributed processing provides an effective framework for improving diabetes classification accuracy while reducing dimensionality and computational burden.

***Keywords:*** *Diabetes diagnosis; Group Learning Optimization; feature selection; Generative Adversarial Network; Apache Spark; Random Forest; machine learning; distributed processing*

## 1. Introduction

Diabetes mellitus is one of the most consequential chronic metabolic disorders in contemporary healthcare because of its increasing prevalence, prolonged clinical course, heterogeneous manifestations, and substantial burden on patients and health systems. The disease is characterized principally by persistent disturbances in glucose regulation, but its consequences extend well beyond glycemic control and may involve cardiovascular, renal, neurological, ophthalmological, and metabolic complications. Population-based evidence has also demonstrated that type 2 diabetes is closely associated with modifiable lifestyle and metabolic risk factors, emphasizing the importance of identifying high-risk individuals before irreversible complications occur (1). Biological studies have further shown that the onset of diabetes may be accompanied by alterations in physiological processes such as renal steroidogenesis, illustrating the systemic nature of the disease and the complexity of its underlying mechanisms (2). Moreover, the diagnosis itself frequently requires substantial lifestyle adaptation by patients, including changes in nutrition, physical activity, self-monitoring, and other health behaviors, which means that delayed or inaccurate diagnosis can affect both medical outcomes and subsequent self-management (3). These characteristics collectively support the need for diagnostic systems capable of identifying diabetes accurately, efficiently, and as early as possible.

Accurate diabetes classification remains challenging because clinically relevant variables interact in nonlinear and heterogeneous ways, while disease manifestations may vary across individuals, populations, and stages of progression. Inaccurate classification is particularly consequential when the type or status of diabetes is incorrectly identified, since such errors may influence therapeutic decisions and long-term disease management (4). Conventional screening and diagnostic approaches typically rely on combinations of biochemical measurements, clinical history, anthropometric indices, and physician interpretation. Although such approaches are indispensable, data-driven prediction models can complement them by detecting multivariate patterns that may not be obvious through isolated clinical measures. Recent work has consequently emphasized the application of machine learning to the early detection of chronic diseases, with diabetes serving as a prominent case because sufficiently accurate prediction may facilitate earlier intervention and risk reduction (5). Contemporary comparative evaluations using benchmark diabetes datasets have similarly demonstrated that machine-learning models can extract clinically meaningful relationships from routine health variables and provide useful predictive information (6).

Machine learning has therefore become increasingly prominent in diabetes research. Artificial neural networks, support vector machines, decision trees, ensemble models, deep learning architectures, and hybrid intelligent systems have all been investigated for the prediction or classification of diabetes. Artificial neural network approaches have demonstrated the ability to model nonlinear relationships among diabetes-related features and produce useful predictive performance (7). Support vector machines have also shown considerable promise in medical classification because their margin-based formulation can provide effective separation between classes even in complex feature spaces. For example, nonlinear heart-rate variability features have been used with artificial neural networks and support vector machines to distinguish diabetic states, indicating that machine-learning methods can exploit physiological information beyond conventional laboratory measurements (8). Decision-tree approaches have likewise been successfully used in high-risk adult populations, offering the additional advantage of comparatively interpretable decision structures (9). Comparative studies involving classifier and hybrid machine-learning techniques further reinforce the value of algorithmic approaches for improving diabetes prediction (10).

More recent research has extended diabetes prediction toward increasingly sophisticated architectures. Deep learning has been employed to learn hierarchical representations of diabetes-related data without depending exclusively on manually constructed decision rules. Convolutional long short-term memory architectures have been investigated for diabetes detection and have demonstrated the potential of sequential and deep representation-learning strategies (11). Bidirectional long short-term memory models have also been proposed to improve the extraction of relationships within diabetes data



Health Nexus
E-ISSN: 2981-2569

(12). Another innovative approach converts structured clinical information into image representations so that deep neural architectures can be applied to diabetes prediction (13). Deep-learning frameworks have similarly been developed specifically for diabetes diagnosis and have shown that learned representations may improve diagnostic discrimination when adequate data and training resources are available (14). In parallel, deep unsupervised approaches combining ensemble feature selection and deep belief neural networks have been explored for early diabetes-risk prediction, illustrating the continued convergence of representation learning and feature optimization (15).

Despite this progress, high predictive complexity does not automatically guarantee clinically useful or computationally efficient models. Diabetes datasets frequently contain redundant, weakly informative, or correlated attributes, and the presence of unnecessary features can increase training time, computational cost, and the risk of overfitting. The influence of individual health-related attributes on diabetes prediction has therefore become an important research issue, with evidence indicating that careful analysis of feature relevance can meaningfully affect model performance (16). Risk-scoring research similarly demonstrates that diabetes prediction can be strengthened by identifying combinations of variables that contribute most effectively to individual risk estimation (17). Accordingly, feature selection is not merely a computational convenience; it is a central component of predictive modeling because it seeks to preserve the most discriminative information while excluding variables that make little contribution to the classification task.

Feature-selection methods have been extensively studied in medical machine learning because the optimal subset of predictors is rarely known in advance. Conventional dimensionality-reduction and optimization approaches have shown that eliminating irrelevant or redundant variables can substantially alter diabetes-classification performance. The combination of principal component analysis and particle swarm optimization, for example, has been investigated for improving diabetes classification (18), while other work has demonstrated that dedicated feature-selection and dimensionality-reduction strategies can enhance the effectiveness of machine-learning algorithms for diabetes prediction (19). Hybrid feature-selection procedures have also been specifically designed to increase diabetes diagnostic accuracy by combining complementary search mechanisms (20). These findings suggest that the quality of the selected subset may be as important as the choice of the final classifier.

Metaheuristic optimization provides a particularly attractive framework for feature selection because selecting an optimal subset from a large candidate space is fundamentally a combinatorial optimization problem. Rather than exhaustively evaluating every possible subset, metaheuristic algorithms search the solution space using population-based, evolutionary, or biologically inspired mechanisms. Genetic algorithms have been applied to feature selection and parameter optimization in cardiovascular prediction, illustrating their capacity to jointly optimize model structure and predictor subsets (21). Binary Crow Search has similarly been developed for feature-selection problems through mechanisms adapted to discrete decision spaces (22). Moth Flame Optimization has been adapted for intelligent feature selection in medical diagnosis (23), while improved Grasshopper Optimization has been used for simultaneous hyperparameter estimation and feature selection (24). Such approaches indicate that swarm and evolutionary intelligence can provide flexible search strategies for medical feature-selection problems.

A major limitation of metaheuristic methods, however, is their susceptibility to an inappropriate balance between exploration and exploitation. Excessive exploration may prevent convergence, whereas premature exploitation may trap the search process around locally optimal feature subsets. This problem has motivated the development of increasingly sophisticated optimization algorithms. Harris Hawks Optimization was introduced as a population-based method inspired by cooperative hunting behavior and has demonstrated strong global optimization capabilities (25). An improved Harris Hawks Optimization algorithm combined with simulated annealing has subsequently been applied to medical feature selection to enhance search effectiveness (26). The Water Strider Algorithm represents another metaheuristic designed to improve the exploration of complex optimization landscapes through biologically inspired search mechanisms (27). Bio-inspired machine-learning approaches have also been applied directly to type 2 diabetes detection, emphasizing the growing role of


Health Nexus
E-ISSN: 2981-2569

intelligent optimization in diabetes classification (28). These developments provide a strong rationale for designing optimization strategies capable of adapting their search behavior over time and exploiting information from multiple high-quality candidate solutions.

The present study addresses this issue through Group Learning Optimization, which conceptualizes optimization as a collective learning process among individuals with different levels of competency. This general principle is consistent with the broader development of intelligent optimization methods in which solution quality determines how information is propagated throughout the population. Instead of relying only on a single best individual, a group-learning mechanism can potentially exploit information from several strong candidate solutions while progressively eliminating inferior solutions. This design is especially relevant to feature selection because multiple promising subsets may contain complementary information, and premature convergence around one subset may prevent the discovery of more parsimonious or diagnostically informative combinations. The use of a binary GLO formulation therefore provides a mechanism for mapping continuous search operations into discrete feature-selection decisions while retaining the adaptive search behavior of the original optimization framework.

Another major challenge in diabetes prediction concerns class imbalance. Clinical datasets often contain substantially different numbers of diseased and non-diseased observations. When conventional classifiers are trained directly on such data, they may achieve apparently acceptable overall accuracy while failing to adequately detect the minority class. This is particularly problematic in disease prediction, where false negatives can have serious consequences. Recent diabetes research has therefore emphasized imbalance-handling techniques as an integral component of predictive modeling, showing that appropriately balancing class distributions can improve machine-learning performance (29). Generative Adversarial Networks provide a particularly powerful solution because they can learn the distribution of minority-class observations and generate synthetic samples that are more flexible than simple duplication or interpolation strategies. Hybrid GAN-based approaches have already demonstrated effectiveness for resolving imbalanced-data problems in other predictive settings (30), providing a methodological basis for extending GAN-based balancing to diabetes datasets.

The use of GANs is especially attractive when minority-class observations must be increased without merely replicating existing cases. Through adversarial interaction between a generator and discriminator, GANs can progressively produce artificial observations that approximate the underlying distribution of real samples. When appropriately integrated with feature selection, this procedure may improve both the representativeness of training data and the reliability of subsequent feature-subset evaluation. This is important because an optimization algorithm may otherwise identify feature combinations that primarily explain the majority class rather than the clinically important minority class. Balancing therefore needs to occur within a carefully structured analytical pipeline rather than being treated as an isolated preprocessing operation.

Diabetes prediction also increasingly occurs in environments characterized by large and heterogeneous healthcare data. Electronic health records, hospital information systems, wearable technologies, mobile monitoring, Internet of Things devices, and connected clinical infrastructures have greatly increased the volume and velocity of data available for analysis. Big-data analytics consequently offers substantial opportunities for healthcare enhancement, while simultaneously introducing challenges related to storage, processing, integration, privacy, and computational scalability (31). Healthcare 4.0 frameworks explicitly integrate cloud computing, fog computing, IoT, big data, and machine learning to enable more responsive and intelligent health services (32). Emerging 5G infrastructures may further strengthen these systems by providing higher-speed communication and reduced latency, although their implementation also introduces broader technical and infrastructural challenges (33). Diabetes monitoring systems have already been proposed using deep belief neural networks, edge IoT, big-data processing, and 5G-supported communication, demonstrating the feasibility of integrating prediction with distributed digital-health environments (34).

The growth of healthcare datasets creates a corresponding need for distributed machine-learning platforms. Apache Spark has become a major framework for scalable analytics because it supports in-memory distributed computation and



Health Nexus
E-ISSN: 2981-2569

can execute machine-learning operations across multiple nodes. Research on large-scale classification has demonstrated the suitability of Spark for distributing computational workloads and improving the processing of high-volume datasets (35). Optimization-enabled Spark frameworks have likewise been proposed for handling large data streams, supporting the integration of intelligent optimization with distributed computation (36). Spark has also been used in healthcare-oriented privacy-preserving big-data architectures, illustrating its applicability beyond conventional analytical tasks (37). In related medical applications, machine-learning analysis on Spark has successfully processed large-scale health information for disease identification (38).

Distributed optimization is particularly relevant to feature selection because searching large feature spaces and repeatedly evaluating candidate subsets may become computationally expensive as dataset size increases. Spark-based distributed Particle Swarm Optimization has been used for feature selection and disease prognosis, demonstrating that parallel execution can substantially improve the scalability of optimization-driven medical classification (39). Similarly, MapReduce-based optimized gradient-boosted-tree frameworks have been investigated for diabetes diagnosis, indicating that distributed architectures can support both predictive performance and large-scale computational efficiency (40). These findings suggest that integrating feature-selection algorithms with Apache Spark can address a limitation that is often overlooked in conventional diabetes-prediction studies: a method may demonstrate high classification accuracy on a small benchmark dataset but remain unsuitable for realistic hospital-scale data if its computation cannot be efficiently distributed.

The broader diabetes literature also illustrates the diversity of prediction problems that machine-learning systems must address. Models have been developed to estimate the risk of hospital readmission among diabetic patients, demonstrating the potential of computational approaches for supporting post-diagnosis clinical management (41). Machine-learning techniques have been assessed for type 2 diabetes risk prediction across alternative model configurations (42), while Hidden Naïve Bayes has been comparatively evaluated as another probabilistic approach to diabetes prediction (43). Machine-learning models have also been used to predict gestational diabetes through combinations of deep learning, Bayesian optimization, and conventional classifiers (44). Recent work combining machine learning with explainable artificial intelligence further emphasizes that predictive accuracy should ideally be accompanied by information about the variables and mechanisms contributing to diabetes predictions (45). Collectively, these studies demonstrate that no single algorithmic family is universally optimal and that diabetes prediction benefits from combining complementary methodological components.

Related research on diabetes complications reinforces the same conclusion. Machine-learning and optimization techniques have been applied to diabetic retinopathy using support vector machines combined with Biogeography-Based Optimization (46). Hybrid ensemble feature-selection, segmentation, and deep majority-voting frameworks have been proposed for large multiclass diabetic-retinopathy databases (47), while deep convolutional neural networks have been used to detect retinopathy through the extraction of both local and long-range image dependencies (48). Although these studies address a complication rather than primary diabetes diagnosis, they demonstrate the value of combining feature optimization, ensemble decisions, and advanced learning architectures in diabetes-related medical analytics. Likewise, evidence concerning possible therapeutic effects of interventions such as curcumin on glycemic and lipid profiles underlines the biological and clinical complexity of diabetes and the importance of accurately characterizing patients using multiple relevant indicators (49).

Recent predictive models increasingly combine several intelligent techniques rather than relying on a single classifier. Evolutionary ensemble prediction based on Harmony Search and stacking has been proposed for diabetes diagnosis, reflecting the growing emphasis on optimization-assisted ensembles (50). Machine-learning studies based on established diabetes datasets have similarly highlighted the value of comparing models across different populations and datasets rather than assuming that a single configuration will generalize uniformly (6). Hybrid and bio-inspired methods continue to be explored because they can simultaneously address feature relevance, classifier


Health Nexus
E-ISSN: 2981-2569

performance, hyperparameter tuning, and search complexity (51). This methodological trajectory suggests that future improvements are likely to arise from integrated architectures in which data preparation, optimization, classification, and computational infrastructure are jointly designed.

Despite extensive prior research, an important gap remains. Many diabetes-prediction studies focus primarily on improving classifier accuracy without simultaneously addressing class imbalance, feature redundancy, metaheuristic convergence behavior, and distributed computational scalability. Other studies optimize feature subsets but evaluate them only in centralized environments, while distributed-processing studies may not incorporate an adaptive feature-selection stage before task distribution. Furthermore, an optimization algorithm that performs adequately on a single diabetes dataset should ideally demonstrate broader numerical robustness on recognized benchmark functions before its medical feature-selection performance is interpreted. These limitations indicate the need for a unified framework capable of balancing the dataset, identifying an informative and parsimonious feature subset, maintaining strong global-search performance, and distributing the reduced computational workload across scalable processing nodes.

Accordingly, the present study integrates GAN-based class balancing, a binary Group Learning Optimization algorithm, and Apache Spark distributed processing with Support Vector Machine, Decision Tree, Random Forest, and majority-voting classification to construct a scalable diabetes-diagnosis framework. The proposed design is intended to improve several interconnected aspects of computational diagnosis: GAN increases minority-class representation; GLO searches adaptively for informative feature subsets; binary transfer functions convert optimization solutions into explicit feature-selection decisions; and Apache Spark distributes reduced-dimensional data across worker nodes for parallel learning. Such integration is expected to improve diagnostic accuracy while simultaneously controlling dimensionality and execution time. The design is further evaluated both on standard optimization benchmark functions and on diabetes-related datasets so that the search properties of GLO can be distinguished from its downstream predictive contribution.

The aim of this study was to develop and evaluate a diabetes diagnosis approach that combines GAN-based data balancing, binary GLO-based feature selection, and Apache Spark distributed processing to improve feature parsimony, classification performance, and computational efficiency.

## 2. Methods and Materials

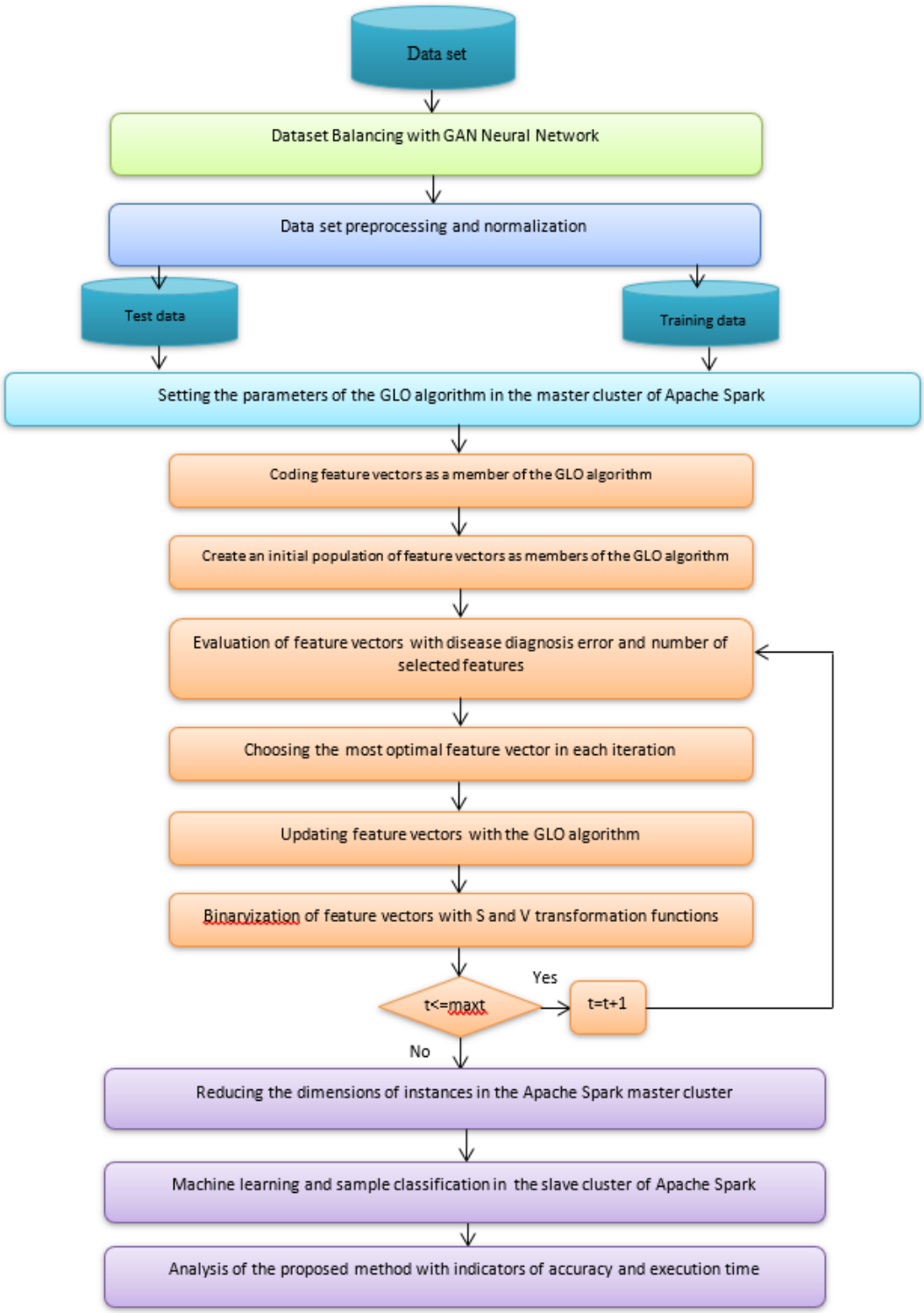
The present study was designed as a computational and predictive diagnostic investigation aimed at developing an efficient diabetes classification framework through the integration of data balancing, metaheuristic feature selection, machine-learning classification, and distributed data processing. The proposed approach combined a Generative Adversarial Network (GAN) for correcting class imbalance, a binary Group Learning Optimization (GLO) algorithm for identifying the most informative diagnostic variables, and Apache Spark distributed processing technology for implementing scalable classification. The analytical workflow was organized sequentially so that diabetes-related data were initially collected and preprocessed, the imbalance between diagnostic classes was corrected using GAN-generated synthetic observations, the balanced dataset was then transferred to the Apache Spark master node, and the most relevant diagnostic features were selected through the binary GLO algorithm. After dimensionality reduction, the reduced dataset was distributed among Spark worker nodes, where Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) algorithms were applied for diabetes classification. This architecture was intended to reduce unnecessary computational complexity while maintaining or improving diagnostic accuracy.



Health Nexus
E-ISSN: 2981-2569

**Figure 1**

*Framework and stages of the proposed diabetes diagnosis method*




The study used the Pima Indians Diabetes Dataset (PIDD), which is one of the most frequently employed benchmark datasets for the development and evaluation of machine-learning methods for diabetes diagnosis. The dataset consists of 768 observations and includes eight predictor variables together with a binary diagnostic outcome. The predictor variables include number of pregnancies, plasma glucose concentration, diastolic blood



Health Nexus
E-ISSN: 2981-2569

pressure, triceps skinfold thickness, serum insulin concentration, body mass index, diabetes pedigree function, and age. The outcome variable identifies whether an individual belongs to the diabetic or non-diabetic class. Each observation was therefore represented as a numerical feature vector associated with a binary diagnostic label. Because the investigation relied on an existing anonymized benchmark dataset and involved no direct contact with human participants, the study did not involve recruitment, intervention, or collection of personally identifiable clinical information.

The proposed diagnostic process consisted of several interconnected computational stages. First, diabetes-related data were prepared for analysis and the numerical attributes were normalized to provide a comparable numerical scale. Second, class imbalance was addressed through GAN-based synthetic data generation, with particular emphasis on increasing the representation of the minority class. Third, the balanced dataset was provided as input to the master cluster in the Apache Spark environment. Fourth, the binary GLO algorithm searched for an optimal subset of diagnostic variables by simultaneously considering classification error and the number of selected features. Fifth, the dimensionality of the diabetes dataset was reduced according to the feature subset selected by GLO. Finally, the reduced data were distributed among Spark worker nodes and classified through SVM, Decision Tree, and Random Forest models. The predicted diagnostic outcomes were then compared with the actual diabetes labels to evaluate the overall effectiveness of the proposed framework.

The principal data source was the structured PIDD database, in which each observation contained a set of clinical and physiological attributes relevant to diabetes prediction. Before feature selection and classification, preprocessing was performed to increase the reliability of the learning process. Because the original attributes were expressed in different units and numerical ranges, normalization was used to prevent variables with relatively large numerical values from exerting disproportionate influence on the optimization and classification processes. For normalization within an arbitrary interval [a,b], the following transformation was employed:

$$N'(x) = ((N(x) - \min(x)) / (\max(x) - \min(x))) \times (b - a) + a$$

When the data were normalized specifically to the interval [0,1], the transformation was expressed as:

$$N'(x) = (N(x) - \min(x)) / (\max(x) - \min(x))$$

In these equations, N(x) represents the original value of feature x, N'(x) represents its normalized value, min(x) and max(x) indicate the minimum and maximum observed values of that feature, and a and b correspond to the lower and upper limits of the selected normalization interval. The normalized data were subsequently used during the feature-selection and classification stages.

A major component of data preparation was the correction of class imbalance. An imbalanced dataset may cause a machine-learning classifier to preferentially learn the characteristics of the majority class while inadequately recognizing minority-class observations. This problem is especially important in medical diagnosis because insufficient identification of the minority disease class can substantially reduce clinical usefulness. To reduce this problem, the proposed method applied a GAN to generate synthetic observations corresponding to the minority diabetes class. The GAN architecture consisted of two competing neural networks: a generator G and a discriminator D. The generator received random noise z and generated artificial observations intended to resemble actual minority-class samples. The discriminator simultaneously received real observations and generator-produced synthetic observations and attempted to distinguish between them. Through repeated adversarial training, the generator progressively produced samples that more closely approximated the distribution of genuine diabetes observations.

The objective function of the GAN was formulated as:

$$\min_G \max_D F(G,D) = E_{x\sim P_d}[\log D(x)] + E_{z\sim n_z}[\log(1 - D(G(z)))]$$

In this expression, $P_d$ represents the probability distribution of real samples, z denotes the random noise vector, $n_z$ represents the distribution from which the noise samples are generated, G(z) represents an artificial observation produced by the generator, and D(.) indicates the probability estimated by the discriminator that a particular observation is real. The generator attempted to minimize the probability that generated observations were identified as artificial, whereas the discriminator attempted to maximize its ability to distinguish genuine from synthetic

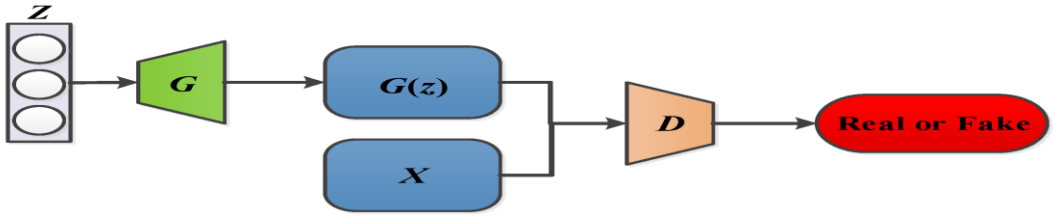


Health Nexus
E-ISSN: 2981-2569

observations. After sufficient adversarial learning, synthetic minority-class examples were added to the original data to increase the number of diabetes-related observations and produce a more balanced dataset.

**Figure 2**

*Structure of the GAN neural network used for dataset balancing*




The primary feature-selection instrument was the Group Learning Optimization algorithm. GLO was developed as an improved human-learning-inspired optimization method based conceptually on Teaching-Learning-Based Optimization. In conventional TLBO, learning is primarily guided by the best teacher, which may limit the contribution of other competent solutions and increase the possibility of premature convergence toward a local optimum. Furthermore, low-quality candidate solutions may remain in the population for repeated iterations even when their contribution to the optimization process is negligible. GLO was designed to address these limitations by modeling group learning as a process in which multiple competent individuals may act as educators, less competent members primarily function as learners, and unsuitable solutions may be competitively eliminated and replaced.

The population of candidate solutions in GLO was defined as:

$H = \{H_1, H_2, ..., H_n\}$

where $H_i$ represents the i-th candidate solution and n denotes the population size. Each candidate solution consisted of D dimensions:

$H_i = \{H_i^1, H_i^2, ..., H_i^D\}$

The initial population was generated randomly within the lower and upper boundaries of the optimization problem according to:

$H_i = l + (u - l) \times rand(0,1)$

where l and u denote the lower and upper limits of the search space. Each candidate solution was evaluated using an objective function $f(H_i)$, after which an educational weight was assigned according to its relative competency:

$w_i = (f(H_i) - \min(f(H))) / (\max(f(H)) - \min(f(H)))$

The resulting weight satisfied:

$0 \leq w_i \leq 1$

The weight represented the relative influence of each population member in the learning process. More competent individuals were therefore permitted to exert a greater educational influence on other candidate solutions. The number of learners potentially influenced by a teacher was expressed as:

$L = w_i \times Max$

where Max denotes the maximum number of learners that a competent individual can guide. The average fitness of the population was calculated as:

$M = (\Sigma_{i=1}^{n} f(H_i)) / n$

The average fitness was used as a criterion for separating competent individuals from weaker solutions. Members performing above the defined competency level could act as teachers, whereas weaker members primarily participated as learners.

The teaching phase modified candidate solutions by moving them toward competent individuals:

$H_i^{new} = Teacher + R(t) \times rand$

The variable R(t) controlled the search radius and gradually decreased over the optimization process:

$R(t) = (R(1) - (R(1) \times t / Max_t)) \times U(t)$

where t is the current iteration, $Max_t$ denotes the maximum number of iterations, and R(1) represents the initial search-radius coefficient. A quasi-random mechanism was introduced through:

$U(t+1) = \cos(t \times \cos^{-1}(U(t)))$

with U(1) initialized to one. This mechanism allowed relatively broad global exploration during the initial iterations and increasingly focused local exploitation as the algorithm approached termination.



Health Nexus
E-ISSN: 2981-2569

Learning also occurred directly between population members. When a weaker solution H_i learned from a more competent solution H_j, its position was updated as:

H_i^new = H_i + rand × (H_j - H_i)

Group-level learning from the mean knowledge of the population was expressed as:

H_i^new = H_i + rand × ((Σ_(i=1)^n H_i / n) - H_i)

Following the teaching and learning stages, competitive elimination was applied. Candidate solutions with relatively weak fitness had a greater probability of being removed, and new solutions were generated to replace them. Through this procedure, poorly performing solutions did not necessarily remain in the population throughout the complete optimization process. The combination of multiple educators, learner interaction, population-level learning, and competitive elimination was intended to provide a better balance between global exploration and local exploitation.

Because the original GLO algorithm operates in a continuous search space, it was converted into a binary form for feature selection. The complete set of available variables was represented as:

A = {F^1, F^2, F^3, ..., F^n}

The feature-selection problem sought an optimal subset S from the complete set A:

S = {F^i | F^i ∈ A, S ⊆ A}

where n denotes the total number of available features and m denotes the number of selected features, with m ≤ n. In the binary GLO representation, every candidate solution corresponded to a binary feature-selection vector. A value of one indicated that the corresponding diabetes-related feature was selected, whereas a value of zero indicated that the feature was excluded. In this manner, each member of the GLO population represented a different candidate subset of diagnostic variables.

The fitness of each candidate feature vector was determined by an objective function that simultaneously considered classification error and the number of retained variables:

Cost(H_i) = α × γ_R(D) + β × (|R| / |N|)

where γ_R(D) denotes the classification error associated with the selected feature subset, |R| represents the number of selected variables, |N| represents the total number of available features, α determines the relative contribution of classification error, and β = 1 - α determines the relative contribution of feature reduction. A multilayer neural network was used as the internal classifier for evaluating the quality of candidate feature subsets. Accordingly, the optimal feature vector was expected to minimize both diagnostic prediction error and the number of selected variables.

Because GLO generates continuous values during its update process, transfer functions were required to convert candidate values into binary feature-selection decisions. Both V-shaped and S-shaped transfer functions were considered. The V-shaped transformation was expressed as:

T(H_i) = |H_i / sqrt(a + H_i^2)|

The S-shaped transfer function was expressed as:

T(H_i) = 1 / (1 + e^(-aH_i))

The outputs of these transfer functions fell within the interval [0,1]. A threshold value of 0.5 was subsequently used to generate the final binary decision:

H_i = 0 if T(H_i) ≤ 0.5

H_i = 1 if T(H_i) > 0.5

Thus, after each GLO updating operation, the candidate feature vectors were transformed into binary masks. Each mask was mapped onto the diabetes dataset, and only the variables corresponding to binary values equal to one were retained for evaluation. The selected variables were supplied to the multilayer neural network, and the resulting classification error and number of retained variables were used to calculate the cost of the candidate solution. The feature vector achieving the smallest objective-function value was retained as the optimal solution for the corresponding iteration, and the procedure was repeated until the maximum number of iterations was reached.

The dataset was divided into 70% training observations and 30% testing observations. The training data were used to guide the optimization and classifier-learning procedures, whereas the testing data were retained for assessing the predictive performance of the final diagnostic models. The optimal binary feature vector obtained from GLO was applied as a dimensionality-reduction mask so that only the most informative diabetes-related attributes were retained. The resulting reduced-dimensional dataset was then prepared for distributed processing within the Apache Spark environment.

Data analysis was performed through three major computational stages consisting of GAN-based class



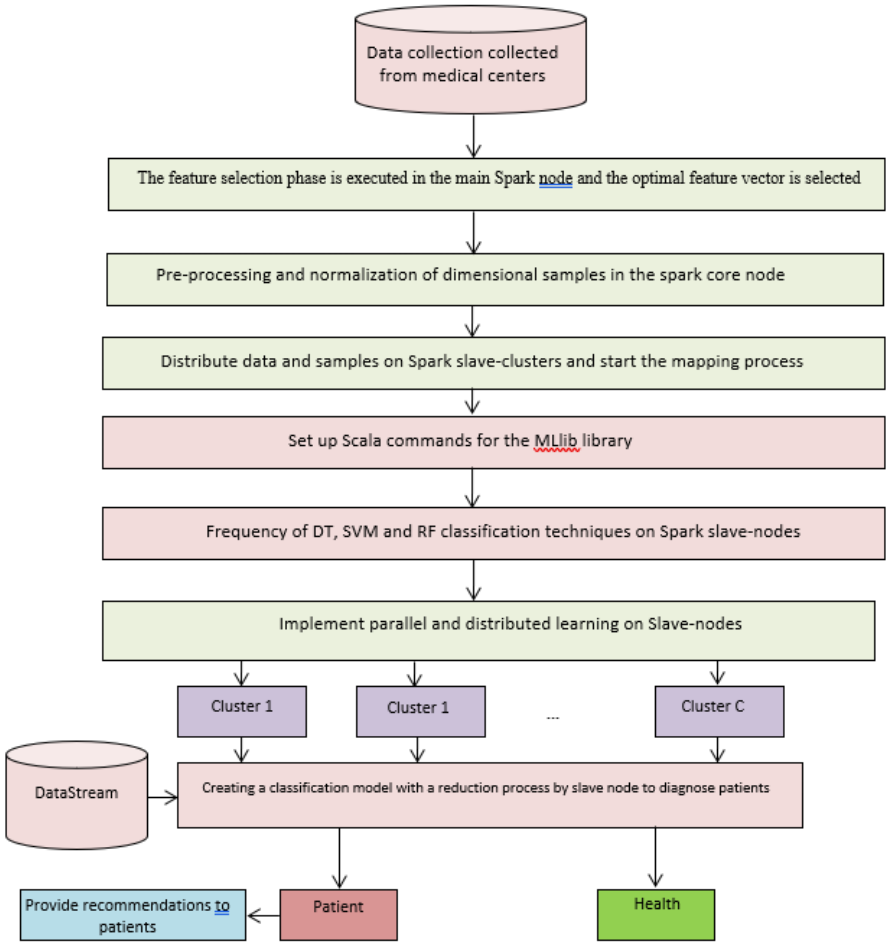
Health Nexus
E-ISSN: 2981-2569

balancing, GLO-based feature selection, and Apache Spark-based distributed classification. Initially, the normalized diabetes dataset was processed using the GAN to increase the representation of the minority class. The balanced dataset was then transferred to the Spark master node, where the binary GLO algorithm was applied. An initial population of randomly generated feature vectors was created, and each feature vector was mapped onto the dataset. The corresponding subset of variables was evaluated using the multilayer neural-network-based objective function. Competency weights were calculated for individual candidate solutions, qualified solutions participated as teachers, weaker solutions learned from competent individuals and from the population mean, and low-quality solutions were subjected to competitive elimination. The candidate vectors were transformed into binary representations using the selected transfer function during each iteration. At the final iteration, the feature vector associated with the minimum cost was selected and used to reduce the dimensionality of the diabetes dataset.

After feature selection, the reduced-dimensional data were processed within the Apache Spark distributed architecture. The master node was responsible for coordinating data preparation, feature reduction, and task allocation, whereas Spark worker nodes processed distributed partitions of the reduced dataset. Performing feature selection before distributed classification ensured that only relevant variables were transferred to the worker nodes, thereby reducing computational overhead, memory consumption, and unnecessary data communication. The reduced data were then classified using Support Vector Machine, Decision Tree, and Random Forest models implemented across the Spark processing environment.

**Figure 3**

*Distributed diabetes diagnostic model using GLO feature selection and Apache Spark*





Health Nexus
E-ISSN: 2981-2569

The Support Vector Machine classifier was applied to construct a decision boundary that maximally separated diabetic and non-diabetic observations. For a linear decision function, the classification model was represented as:

$f(x) = w^T x + b$

where x represents the input feature vector, w denotes the weight vector, and b is the bias term. The predicted class was determined according to the sign of the resulting decision function. SVM was included because margin maximization can provide effective generalization when the number of available observations is limited relative to the complexity of the feature space.

The Decision Tree classifier modeled diabetes status through a sequence of recursive feature-based partitions. At each internal node, a selected predictor was used to divide observations into increasingly homogeneous groups. The recursive splitting process continued until a stopping criterion was reached, and terminal nodes were assigned diagnostic classes. Decision Tree classification was included because its hierarchical decision rules provide a comparatively interpretable representation of relationships between selected clinical predictors and diabetes status.

Random Forest classification was performed as an ensemble extension of the Decision Tree approach. Multiple decision trees were constructed using bootstrap samples of the training observations and randomly selected subsets of candidate predictors. Each tree independently generated a diabetes-status prediction, and the final classification was obtained through aggregation of the individual tree outputs. The ensemble structure was expected to reduce model variance and decrease the tendency toward overfitting that may occur when a single Decision Tree is used.

The distributed Spark implementation allowed partitions of the reduced dataset to be processed concurrently by different worker nodes. The master node coordinated task allocation, data distribution, and integration of classification outputs, while worker nodes executed the assigned machine-learning operations on their respective partitions. This configuration was selected to improve computational scalability and enable the diagnostic framework to process increasingly large diabetes-related datasets without requiring all classification computations to be executed sequentially on a single processing unit.

The diagnostic performance of the proposed method was evaluated by comparing predicted diabetes labels with actual class labels. Four primary outcomes were derived from the resulting confusion matrix. True positives represented diabetic observations correctly classified as diabetic, true negatives represented non-diabetic observations correctly classified as non-diabetic, false positives represented non-diabetic observations incorrectly classified as diabetic, and false negatives represented diabetic observations incorrectly classified as non-diabetic. These values were used to calculate accuracy, precision, sensitivity, specificity, and F1-score.

Accuracy was computed as:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Precision was computed as:

$$\text{Precision} = TP / (TP + FP)$$

Sensitivity was computed as:

$$\text{Sensitivity} = TP / (TP + FN)$$

Specificity was computed as:

$$\text{Specificity} = TN / (TN + FP)$$

The F1-score was calculated as:

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Sensitivity}) / (\text{Precision} + \text{Sensitivity})$$

These measures were evaluated jointly rather than relying exclusively on accuracy. Sensitivity was particularly important because it quantified the ability of the diagnostic model to correctly detect individuals with diabetes, whereas specificity reflected the ability to correctly recognize non-diabetic observations. Precision indicated the proportion of positive diagnostic predictions that were correct, while the F1-score summarized the balance between precision and sensitivity.

The final analytical comparison focused on the diagnostic performance of SVM, Decision Tree, and Random Forest after GAN-based class balancing and binary-GLO feature selection. The number of selected features was also considered because the GLO objective function explicitly penalized unnecessarily large feature subsets. Consequently, the most desirable result was defined as a model capable of achieving high classification performance while relying on a comparatively compact subset of diabetes-related attributes. The complete framework therefore treated diabetes diagnosis as an integrated optimization problem in which class balance, feature relevance, predictive accuracy,



Health Nexus
E-ISSN: 2981-2569

dimensionality reduction, and distributed computational efficiency were jointly considered.

## 3. Findings and Results

The experimental evaluation was conducted in several stages to examine the optimization capability of the proposed Group Learning Optimization (GLO) algorithm, its effectiveness for feature selection in diabetes diagnosis, and its performance when integrated with Apache Spark distributed processing. The first stage evaluated the optimization behavior of GLO using benchmark functions from the Congress on Evolutionary Computation framework. These functions were used as objective or cost functions for assessing the ability of metaheuristic algorithms to approach known global optima. Both unimodal and multimodal benchmark functions were considered. Unimodal functions primarily assess exploitation and local search ability because they generally contain a single global optimum, whereas multimodal functions contain multiple local optima and therefore provide a more demanding assessment of global exploration. For most of the benchmark functions examined in the three-dimensional setting, the global optimum was located at x = 0 and y = 0, with an optimal objective value of f*(0,0) = 0. Functions such as Griewank and Ackley were particularly useful for evaluating global search behavior because their search spaces contain numerous local optima that can trap insufficiently exploratory metaheuristic algorithms.

For implementation of the optimization experiments, the population size was fixed at 15 candidate solutions, the maximum number of iterations was 100, and every experiment was independently repeated 50 times. In the proposed GLO algorithm, U(1) was initialized to 1, while R(t) varied within the interval [0,1]. The GLO algorithm was compared with Teaching-Learning-Based Optimization (TLBO), Whale Optimization Algorithm (WOA), Spotted Hyena Optimizer (SHO), Water Strider Algorithm (WSA), and Harris Hawks Optimization (HHO). Parameter settings for the competing algorithms followed their original recommended configurations. In WOA, coefficient C varied randomly within [0,2], while the initial value of a was set to 2. For SHO, parameter h was set to 5. For HHO, the initial coefficients E and J were both set to 2. In WSA, the coefficient controlling the generation of new solutions was set to 5. All input features were normalized within the interval [0,1] before model implementation.

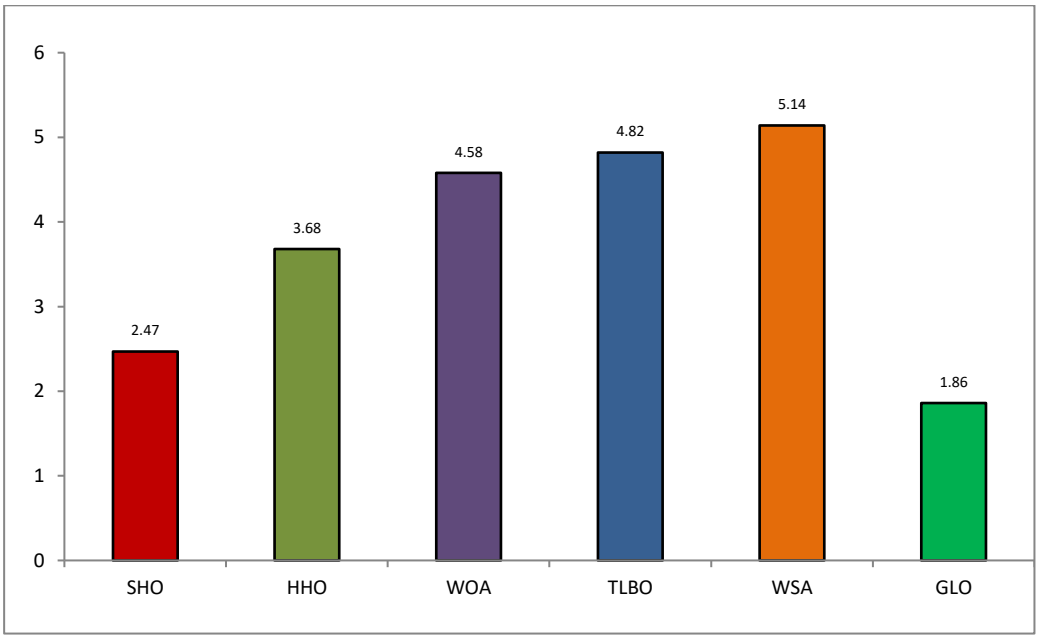
The optimal-solution error, convergence behavior, local-optimum entrapment rate, and comparative algorithm ranking were used to evaluate optimization performance. Examination of the convergence patterns on benchmark functions such as Sphere and Ackley demonstrated that the optimization error of all algorithms generally decreased with increasing iterations. However, GLO showed a more pronounced decline in error and achieved a smaller final optimization error than the competing methods. This pattern indicated that GLO approached the known global solutions with greater numerical precision. The progressive reduction of the search radius in GLO contributed to broad exploration during the early iterations and more intensive exploitation around promising candidate solutions during later iterations.

The aggregate ranking across the benchmark functions further supported the superior optimization performance of GLO. The proposed algorithm obtained the best mean rank of 1.86. SHO ranked second with a mean rank of 2.47, followed by HHO with 3.68. The corresponding rankings for WOA, TLBO, and WSA were 4.58, 4.82, and 5.14, respectively. Because lower ranks indicated greater ability to minimize the optimization error, GLO demonstrated the strongest overall performance among the six evaluated metaheuristic algorithms.


Health Nexus
E-ISSN: 2981-2569

**Figure 4**

*Comparative ranking of the algorithms in finding the optimal solution with minimum error*




An additional analysis examined the probability of convergence to local optima. For this purpose, the proposed and competing algorithms were executed 200 times on benchmark functions characterized by multiple local solutions. GLO demonstrated a local-optimum entrapment rate of only 7.5%, compared with 9.5% for SHO, 10.5% for WOA, 12% for HHO, 13.5% for TLBO, and 17% for WSA. Thus, the proposed algorithm had the lowest probability of becoming trapped in local optima. This finding indicated stronger global exploration and greater robustness against premature convergence. The superior behavior of GLO was attributable to the gradual transition from global to local search, competency-based weighting of population members, learning between individuals, learning from collective population knowledge, and competitive elimination of low-quality solutions.

The comparison with established swarm-intelligence methods further demonstrated that GLO provided a comparatively low optimal-calculation error. WOA, HHO, and WSA had previously demonstrated improvements over conventional evolutionary methods such as Differential Evolution, Genetic Algorithm, and Particle Swarm Optimization in comparable optimization contexts. Because GLO generated lower optimization errors than WOA, HHO, and WSA in the present experiments, the findings indicated that the proposed approach offered a highly competitive optimization mechanism. Unlike methods characterized by rapid early convergence, GLO maintained greater search diversity during the initial phases of optimization and therefore reduced the risk that the population would converge prematurely around suboptimal regions.

The second experimental stage examined the effectiveness of GLO for feature selection in diabetes diagnosis using the Pima Indians Diabetes Dataset. The dataset comprised 768 observations characterized by eight predictor features. Among these observations, 268 belonged to the diabetic class and 500 to the non-diabetic class. The initial class distribution therefore demonstrated substantial imbalance. The diagnostic attributes incorporated information related to pregnancy history, oral glucose tolerance, blood pressure, skinfold thickness, insulin, body mass index, diabetes pedigree characteristics, and age.

The feature-selection experiment used 15 candidate feature vectors and 100 GLO iterations. A multilayer perceptron neural network containing 20 neurons in each layer was used as the internal classifier for evaluating the fitness of candidate feature subsets. Diagnostic performance was assessed using accuracy, sensitivity, and precision. These measures were calculated as:

Accuracy = (TP + TN) / (TP + TN + FP + FN)



Health Nexus
E-ISSN: 2981-2569

Sensitivity = TP / (TP + FN)

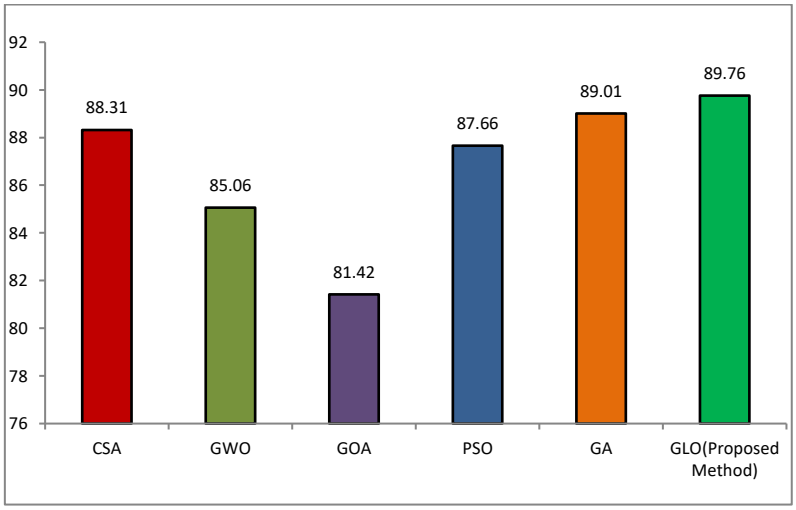
Precision = TP / (TP + FP)

Here, TP represented diabetic cases correctly identified as diabetic, TN represented healthy cases correctly identified as healthy, FP represented healthy cases incorrectly classified as diabetic, and FN represented diabetic cases incorrectly classified as healthy.

When GLO feature selection was evaluated using the multilayer neural network without GAN-based balancing, the proposed method achieved an accuracy of 89.76%, sensitivity of 88.60%, and precision of 88.41%. Its feature-selection accuracy was also compared with several established optimization approaches. The accuracy values obtained using Cuckoo Search Algorithm (CSA), Grey Wolf Optimizer (GWO), Grasshopper Optimization Algorithm (GOA), Particle Swarm Optimization (PSO), and Genetic Algorithm (GA) were 88.31%, 85.06%, 81.42%, 87.66%, and 89.01%, respectively, whereas GLO achieved 89.76%. The proposed method therefore produced the highest diagnostic accuracy among the evaluated feature-selection algorithms. Across repeated experiments, the variables most frequently retained as informative features included gender-related information, oral glucose tolerance, the difference between insulin measurements before fasting and two hours after fasting, and body mass index. The superior classification accuracy obtained after GLO optimization indicated that the algorithm identified a more discriminative feature subset for distinguishing diabetic from non-diabetic observations.

**Figure 5**

*Comparison of diabetes diagnostic accuracy obtained by GLO and competing feature-selection algorithms*




The next analysis investigated whether GAN-based class balancing improved classification after GLO feature selection. Two analytical scenarios were therefore compared. In the first scenario, the original imbalanced PIDD data were used. In the second, the minority class was augmented using GAN-based synthetic sample generation before GLO feature selection and final classification. The optimal feature subset generated in the Spark master node was subsequently evaluated using Decision Tree, Random Forest, Support Vector Machine, and a majority-voting ensemble. The accuracy results are presented in Table 1.

**Table 1**

*Accuracy of the proposed classifiers with imbalanced and GAN-balanced datasets*

| Method | Imbalanced dataset (%) | Balanced dataset (%) |
|---|---|---|
| Decision Tree | 86.26 | 93.21 |
| Random Forest | 87.53 | 94.62 |
| Support Vector Machine | 85.24 | 91.63 |
| Majority voting | 92.39 | 93.92 |



Health Nexus
E-ISSN: 2981-2569

As shown in Table 1, GAN-based balancing improved the accuracy of all three individual classifiers. Decision Tree accuracy increased from 86.26% to 93.21%, representing an absolute improvement of 6.95 percentage points. Random Forest accuracy increased from 87.53% to 94.62%, an improvement of 7.09 percentage points, while SVM accuracy increased from 85.24% to 91.63%, an improvement of 6.39 percentage points. The majority-voting approach also improved from 92.39% to 93.92%. Among the individual classifiers, Random Forest achieved the highest balanced-data accuracy of 94.62%, whereas majority voting produced an accuracy of 93.92%.

Sensitivity was similarly improved after GAN balancing, as shown in Table 2. The Decision Tree sensitivity increased from 85.31% to 91.66%, Random Forest increased from 86.62% to 92.64%, and SVM increased from 84.41% to 90.06%. The majority-voting method achieved the highest sensitivity after balancing, reaching 93.67%, compared with 91.28% when the original imbalanced dataset was used.

**Table 2**

*Sensitivity of the proposed classifiers with imbalanced and GAN-balanced datasets*

| Method | Imbalanced dataset (%) | Balanced dataset (%) |
|---|---|---|
| Decision Tree | 85.31 | 91.66 |
| Random Forest | 86.62 | 92.64 |
| Support Vector Machine | 84.41 | 90.06 |
| Majority voting | 91.28 | 93.67 |

The same general pattern was observed for precision. As reported in Table 3, Decision Tree precision increased from 85.47% to 91.23%, Random Forest increased from 85.59% to 91.78%, and SVM increased from 84.63% to 90.55%. Majority voting achieved the highest balanced-data precision of 92.84%, compared with 91.33% in the imbalanced condition. Taken together, these results showed that GAN-based balancing consistently improved diagnostic performance across accuracy, sensitivity, and precision. With the majority-voting mechanism applied to the balanced data, the proposed framework obtained accuracy, sensitivity, and precision values of 93.92%, 93.67%, and 92.84%, respectively.

**Table 3**

*Precision of the proposed classifiers with imbalanced and GAN-balanced datasets*

| Method | Imbalanced dataset (%) | Balanced dataset (%) |
|---|---|---|
| Decision Tree | 85.47 | 91.23 |
| Random Forest | 85.59 | 91.78 |
| Support Vector Machine | 84.63 | 90.55 |
| Majority voting | 91.33 | 92.84 |

The proposed balanced-data framework was additionally compared with previously reported diabetes prediction methods based on conventional neural networks, convolutional long short-term memory models, combined ResNet18 and ResNet50 architectures, and logistic regression combined with CFA. The proposed approach achieved an accuracy of 93.92% in this comparison. The observed accuracy was higher than the compared ANN, Conv-LSTM, ResNet18+ResNet50, and LR+CFA approaches, indicating that the combination of data balancing, optimized feature selection, and ensemble classification could achieve competitive diagnostic performance without requiring an exclusively deep-learning-based final classifier.

The final experimental stage examined distributed processing within the Apache Spark architecture using a larger dataset containing information collected from 130 hospitals in the United States. The purpose of this stage was to examine both diagnostic performance and computational acceleration when the proposed method was executed in a



Health Nexus
E-ISSN: 2981-2569

distributed environment. Data preprocessing and feature selection were first conducted in the Spark master node. The reduced-dimensional observations were subsequently partitioned among Spark worker nodes, where Decision Tree, Random Forest, and Support Vector Machine classifiers were executed. Scala was used as the implementation language, while Spark MLlib provided the required machine-learning functionality.

The distributed experiments produced particularly high diagnostic performance. Decision Tree achieved an accuracy of 98.40%, sensitivity of 98.13%, and precision of 97.89%. Random Forest obtained the best overall classification performance, with accuracy of 98.97%, sensitivity of 98.64%, and precision of 98.62%. Support Vector Machine achieved an accuracy of 97.58%, sensitivity of 97.21%, and precision of 97.03%. Therefore, all three distributed classifiers exceeded 97% across the reported diagnostic measures, while Random Forest was the strongest individual method in terms of accuracy, sensitivity, and precision.

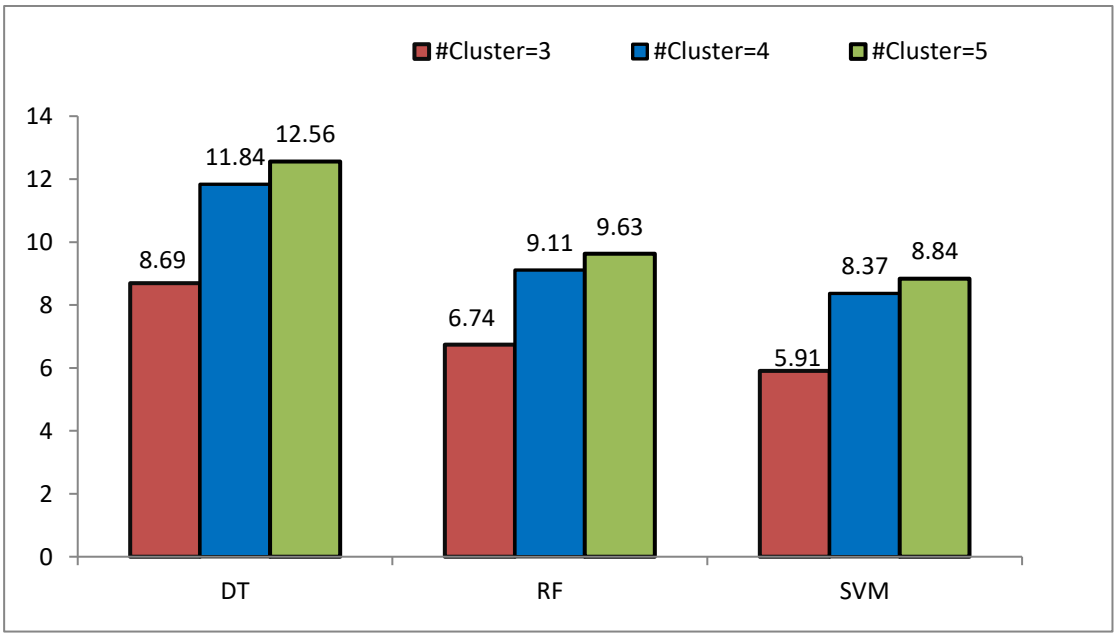
The computational efficiency of the distributed architecture was assessed using the speedup index:

Speedup = TimeRun_non-distributed / TimeRun_distributed

Speedup represented the ratio between the execution time required in a conventional non-distributed environment and the execution time required after distributing the same computation across Apache Spark clusters. Decision Tree, Random Forest, and Support Vector Machine were evaluated using three, four, and five Spark clusters. Each cluster contained a quad-core Intel processor and 6 GB of memory, and the machine-learning procedures were executed using Scala. The experiments demonstrated that computational acceleration increased as the number of Spark clusters increased. This pattern was observed for all three classifiers, indicating that the proposed framework could benefit from increasing distributed computational resources. Decision Tree exhibited the greatest computational acceleration, followed by Random Forest and Support Vector Machine. The maximum recorded speedup was 12.56 for the Decision Tree classifier, demonstrating a substantial reduction in execution time relative to non-distributed processing.

**Figure 6**

*Comparison of computational speedup for Decision Tree, Random Forest, and Support Vector Machine in Apache Spark*




Overall, the experimental findings demonstrated three complementary advantages of the proposed framework. First, GLO showed stronger benchmark optimization performance than the competing metaheuristic algorithms, obtaining the best mean rank of 1.86 and the lowest local-optimum entrapment rate of 7.5%. Second, GLO-based feature selection achieved a diabetes diagnostic accuracy of 89.76% and outperformed CSA, GWO, GOA, PSO, and GA



Health Nexus
E-ISSN: 2981-2569

in the feature-selection comparison. GAN-based class balancing further improved the performance of DT, RF, SVM, and majority-voting classification, with the balanced majority-voting model achieving 93.92% accuracy, 93.67% sensitivity, and 92.84% precision. Third, implementation within Apache Spark produced both strong diagnostic performance and substantial computational acceleration. Random Forest achieved the highest distributed diagnostic accuracy of 98.97%, while Decision Tree achieved the maximum observed speedup of 12.56. These findings supported the effectiveness of combining GAN-based balancing, binary GLO feature selection, and Apache Spark distributed learning for accurate and computationally efficient diabetes diagnosis.

## 4. Discussion and Conclusion

The findings of the present study demonstrated that the proposed diabetes diagnosis framework, which integrated Group Learning Optimization (GLO), GAN-based class balancing, feature selection, and Apache Spark distributed processing, achieved favorable results at both the optimization and diagnostic levels. In the benchmark-function experiments, GLO obtained the best overall mean rank of 1.86, outperforming SHO, HHO, WOA, TLBO, and WSA, whose mean ranks were 2.47, 3.68, 4.58, 4.82, and 5.14, respectively. The proposed algorithm also showed the lowest rate of convergence to local optima, with an entrapment rate of 7.5%, compared with 9.5% for SHO, 10.5% for WOA, 12% for HHO, 13.5% for TLBO, and 17% for WSA. These results suggest that GLO maintained a more effective balance between exploration and exploitation and was less prone to premature convergence. This advantage is consistent with the broader literature on metaheuristic optimization, in which the ability to preserve population diversity while progressively directing the search toward promising regions is regarded as a central determinant of optimization quality. Harris Hawks Optimization, for instance, was developed specifically to improve adaptive exploration and exploitation behavior (25), while the Water Strider Algorithm similarly relies on biologically inspired population interactions to search complex optimization spaces (27). The stronger performance of GLO in the present study indicates that competency-weighted group learning and competitive elimination may provide an effective alternative mechanism for preserving search diversity while intensifying the search around high-quality solutions.

The low local-optimum entrapment observed for GLO can also be interpreted in relation to previous improvements to existing optimization algorithms. Elgamal et al. showed that combining Harris Hawks Optimization with simulated annealing could strengthen feature-selection performance by reducing the probability of premature convergence in medical problems (26). Likewise, intelligent feature-selection approaches based on Moth Flame Optimization have emphasized the importance of balancing local refinement with continued global exploration (23). The GLO mechanism appears to address this same problem through a different behavioral principle. Rather than depending on one dominant leader, multiple sufficiently competent individuals are permitted to influence weaker solutions, while poorly performing members are more likely to be eliminated. This distributed learning structure may explain why the proposed algorithm was able to avoid local optima more frequently than the comparison algorithms. The progressive reduction in the GLO search radius also provided a natural transition from broad exploration during early iterations to more focused exploitation at later stages, which is consistent with the general design principle underlying successful swarm-intelligence algorithms.

The feature-selection phase further confirmed the effectiveness of the proposed optimization strategy in a clinically relevant classification problem. Using the Pima Indians Diabetes Dataset, GLO-based feature selection with a multilayer perceptron classifier achieved 89.76% accuracy, 88.60% sensitivity, and 88.41% precision before GAN-based balancing. The accuracy of GLO exceeded that of CSA, GWO, GOA, PSO, and GA, which achieved 88.31%, 85.06%, 81.42%, 87.66%, and 89.01%, respectively. This finding supports the proposition that effective feature selection can improve diabetes prediction by removing redundant or weakly informative attributes while retaining variables that provide stronger class discrimination. Similar conclusions have been reported in earlier research. Li et al. found that hybrid feature-selection strategies could improve the accuracy of diabetes diagnostic applications by identifying more informative subsets of predictors (20). Sivaranjani et al. likewise demonstrated that combining machine-learning classification with feature


E-ISSN: 2981-2569

selection and dimensionality reduction can enhance diabetes prediction performance (19). Choubey et al. also showed that optimization-based dimensionality reduction, including PSO-related approaches, can improve the performance of diabetes classifiers (18). Taken together, these findings support the present result that model performance depends not only on classifier choice but also on the quality of the diagnostic feature subset supplied to the classifier.

The superiority of GLO over the competing feature-selection methods may be explained by its multi-individual learning mechanism and explicit preference for both lower classification error and smaller feature subsets. The objective function used in the present study penalized unnecessary features while rewarding lower diagnostic error, thereby encouraging parsimonious solutions. Similar multi-objective reasoning has been used in other optimization-based feature-selection studies. Gupta and Sedamkar demonstrated that genetic algorithms can improve predictive learning by jointly optimizing feature selection and model parameters (21), while Chaudhuri and Sahu showed that binary optimization strategies can effectively transform continuous search mechanisms into discrete feature-selection decisions (22). Algamal et al. further emphasized the usefulness of optimization algorithms for simultaneously improving feature selection and predictive-model configuration (24). In the present study, the binary GLO formulation extended this logic by using transfer functions to map continuous search movements into binary inclusion or exclusion decisions while preserving the learning dynamics of the original algorithm.

The variables most frequently retained by the proposed feature-selection procedure were related to glucose tolerance, insulin response, body mass index, and demographic or sex-related characteristics. This pattern is clinically plausible because these variables reflect central metabolic processes associated with diabetes risk and glycemic dysfunction. Previous machine-learning studies have similarly found that the predictive contribution of individual health-related features differs substantially and that glucose-related, anthropometric, and demographic variables often provide strong discriminatory information (16). Risk-modeling research has also shown that combinations of clinical and demographic variables can effectively identify individuals at elevated risk of diabetes (17). The consistency between the features selected by GLO and those highlighted in earlier predictive studies supports the interpretability of the selected feature subset and suggests that the proposed algorithm was not merely optimizing statistical noise.

A particularly important result of the study was the improvement obtained after balancing the PIDD dataset with the GAN. Across all three individual classifiers, balancing increased accuracy, sensitivity, and precision. Decision Tree accuracy increased from 86.26% to 93.21%, Random Forest accuracy from 87.53% to 94.62%, and SVM accuracy from 85.24% to 91.63%. Similar improvements were observed for sensitivity and precision. These findings confirm that class imbalance can materially affect diabetes prediction and that correcting it can improve the recognition of both positive and negative cases. This conclusion is strongly aligned with recent work emphasizing imbalance handling as a central component of reliable diabetes prediction. Abousaber reported that advanced imbalance-management techniques can improve machine-learning-based diabetes prediction (29). Shafqat and Byun similarly demonstrated that GAN-based approaches can effectively address imbalanced-data problems by generating synthetic observations that improve downstream predictive learning (30). The present findings extend this logic to diabetes diagnosis and show that adversarially generated minority-class samples can improve several classifiers when integrated with optimization-based feature selection.

The stronger sensitivity obtained after GAN balancing is particularly important because sensitivity reflects the ability to identify diabetic individuals correctly. In an imbalanced dataset, classifiers may become biased toward the majority non-diabetic class, producing acceptable overall accuracy while missing a substantial proportion of true diabetic cases. By improving minority-class representation, GAN-based balancing appears to have reduced this bias and allowed the classifiers to learn a more representative decision boundary. This is consistent with the broader objective of machine-learning-based diabetes prediction, which is not merely to maximize overall correctness but to identify high-risk or affected individuals sufficiently accurately to support timely intervention (5, 42). The finding is also consistent with evidence that the performance of diabetes classifiers


Health Nexus
E-ISSN: 2981-2569

depends heavily on the characteristics and preparation of the data supplied to them (6).

Among the individual classifiers applied to the balanced PIDD data, Random Forest achieved the highest accuracy at 94.62%, while majority voting produced 93.92% accuracy, 93.67% sensitivity, and 92.84% precision. The strong performance of Random Forest is consistent with the general advantage of ensemble approaches in reducing variance and improving robustness across heterogeneous observations. Previous studies have repeatedly demonstrated the value of hybrid and ensemble learning for diabetes prediction. Barik et al. reported that hybrid machine-learning techniques can improve predictive accuracy compared with individual classifiers (10), while Zhang et al. proposed an evolutionary ensemble model using Harmony Search and stacking for diabetes diagnosis, further illustrating the advantages of combining complementary predictive mechanisms (50). Majority-voting approaches have also shown promise in diabetes-related applications, including large-scale retinopathy classification (47). The present results support the same principle: integrating multiple classifiers can provide a more stable decision than relying on a single model, particularly after feature optimization and class balancing.

The results also showed that the proposed method compared favorably with deep-learning-oriented approaches previously applied to diabetes prediction. Earlier research has demonstrated that convolutional LSTM models (11), bidirectional LSTM architectures (12), deep belief networks (15), and other deep-learning frameworks (13, 14) can achieve strong performance in diabetes prediction. However, these methods frequently require more substantial computational resources, larger datasets, or complex model tuning. The present framework achieved competitive predictive performance through a combination of optimized feature selection, data balancing, and conventional classifiers. This suggests that carefully engineered hybrid systems may provide performance comparable to more computationally intensive deep architectures, particularly when the available clinical dataset is relatively structured and moderate in size.

The distributed Apache Spark experiments provided further support for the proposed architecture. When the framework was applied to data collected from 130 hospitals, Decision Tree achieved 98.40% accuracy, 98.13% sensitivity, and 97.89% precision; Random Forest achieved 98.97%, 98.64%, and 98.62%, respectively; and SVM achieved 97.58%, 97.21%, and 97.03%. Random Forest therefore provided the highest overall diagnostic performance in the distributed setting. These findings are consistent with research showing that distributed data-processing frameworks can support efficient large-scale classification without necessarily sacrificing predictive quality. Jha et al. demonstrated the suitability of Apache Spark for big-data classification (35), while Srivani et al. showed that optimization-enabled Spark architectures can improve the handling of large-scale data streams (36). In healthcare specifically, Spark-based systems have been applied to privacy-preserving big-data processing (37) and disease identification from large-scale patient-related information (38). The present study adds to this literature by combining distributed Spark processing with prior dimensionality reduction through a metaheuristic feature-selection algorithm.

The speedup results provide an additional practical contribution. As the number of Spark clusters increased from three to five, computational acceleration improved for Decision Tree, Random Forest, and SVM, with Decision Tree reaching the maximum observed speedup of 12.56. This result indicates that the architecture can convert additional computational resources into meaningful reductions in execution time. Similar motivations underlie Spark-based distributed feature-selection systems, such as the distributed PSO approach proposed for disease prognosis (39). Selvi and Muthulakshmi also demonstrated the value of distributed MapReduce-oriented machine-learning systems for diabetes diagnosis (40). In the broader context of healthcare big data, such scalability is increasingly important because modern clinical environments generate high-volume data from electronic records, monitoring devices, connected systems, and other digital sources (31, 32). The present results therefore suggest that feature reduction before distributed classification is a useful strategy for limiting unnecessary computational load while preserving strong diagnostic performance.

Taken together, the results indicate that the effectiveness of the proposed framework arose from the interaction of several components rather than from a single algorithmic



Health Nexus
E-ISSN: 2981-2569

element. GLO improved the search for informative feature subsets, GAN balancing strengthened minority-class representation, ensemble and conventional classifiers provided robust predictive learning, and Apache Spark reduced computational time through distributed execution. This integrated design responds to several limitations identified across previous studies, where feature selection, imbalance handling, classifier design, and scalability are often treated separately. The present findings therefore support the value of constructing diabetes-diagnosis systems as coordinated analytical pipelines rather than isolated classification models.

The study nevertheless had several limitations. First, the principal feature-selection experiments relied on the Pima Indians Diabetes Dataset, which contains a relatively limited number of observations and a restricted set of eight predictor variables; therefore, the resulting feature subset may not represent the full diversity of clinical, behavioral, genomic, and sociodemographic factors associated with diabetes in broader populations. Second, the distributed experiments used a separate hospital-based dataset, meaning that direct equivalence between the benchmark diagnostic results and the large-scale Spark results should be interpreted cautiously. Third, the GAN-generated observations were evaluated primarily through downstream classification performance rather than through extensive distributional, clinical, or expert validation of synthetic sample realism. Fourth, the study focused on accuracy, sensitivity, precision, optimization error, and computational speedup, whereas additional measures such as specificity, F1-score, area under the receiver operating characteristic curve, calibration, memory consumption, and communication overhead could provide a more comprehensive assessment. Finally, the proposed framework was evaluated within a specific Spark configuration, and its performance may vary under different hardware, cluster sizes, data-partition strategies, and software environments.

Future research should evaluate the proposed framework using larger, multicenter, geographically diverse, and longitudinal diabetes datasets containing broader clinical and laboratory variables. External validation across independent populations would be particularly useful for determining the generalizability of the selected feature subsets and classification performance. Future studies could also compare GLO with newer binary and multi-objective metaheuristic algorithms and investigate adaptive parameter-control strategies that modify search behavior according to convergence conditions. More extensive evaluation of GAN-generated samples should include distributional similarity, minority-class diversity, clinical plausibility, and resistance to synthetic-data artifacts. In addition, future research could examine alternative balancing approaches, ensemble architectures, explainable artificial intelligence methods, and deep-learning classifiers within the same GLO-Spark framework. Scalability experiments using larger numbers of worker nodes, different Spark partitioning strategies, heterogeneous computing resources, and streaming healthcare data would also help determine the practical limits of the architecture.

For practice, the proposed framework may be most useful as a decision-support architecture rather than as a replacement for clinical diagnosis. Healthcare institutions implementing this approach should establish a standardized pipeline for data cleaning, normalization, class-balance assessment, feature selection, model validation, and distributed execution before deployment. Selected features should be reviewed for clinical plausibility, and model outputs should be presented in a form that healthcare professionals can interpret alongside laboratory findings and patient history. Organizations working with large hospital datasets may benefit from performing dimensionality reduction before distributed classification to decrease computation and communication costs. Routine monitoring of model sensitivity, false-negative rates, dataset drift, and execution time should also be incorporated into implementation so that predictive and computational performance remain stable as new patient data are accumulated.

## Authors' Contributions

All authors contributed substantially to the study and to manuscript development, and all approved the final version.

## Declaration

The authors declare that artificial intelligence tools were used only to assist with language editing, translation, and improvement of the manuscript's readability. All conceptualization, study design, data collection, data



Health Nexus
E-ISSN: 2981-2569

analysis, interpretation of findings, and final approval of the manuscript were performed by the authors. The authors take full responsibility for the accuracy, integrity, and originality of the content.

## Transparency Statement

Data are available for research purposes upon reasonable request to the corresponding author.

## Acknowledgments

The authors thank the participating centers, clinicians, adolescents, and guardians who made the study possible.

## Declaration of Interest

The authors report no conflict of interest.

## Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## Ethics Considerations

The study placed a high emphasis on ethical considerations. Informed consent obtained from all participants, ensuring they are fully aware of the nature of the study and their role in it. Confidentiality strictly maintained, with data anonymized to protect individual privacy. The study adhered to the ethical guidelines for research with human subjects as outlined in the Declaration of Helsinki.

## References

1. Magriplis E, Panagiotakos D, Papakonstantinou E, Mitsopoulou AV, Karageorgou D, Dimakopoulos I, et al. Prevalence Of Type 2 Diabetes Mellitus In A Representative Sample Of Greek Adults And Its Association With Modifiable Risk Factors: Results From The Hellenic National Nutrition And Health Survey. Public Health. 2021. [PMID: 33478772] [DOI]
2. Pagotto MA, Roldán ML, Molinas SM, Raices T, Pisani GB, Pignataro OP, et al. Impairment Of Renal Steroidogenesis At The Onset Of Diabetes. Molecular and Cellular Endocrinology. 2021:111170. [PMID: 33482284] [DOI]
3. Şahin A, Soylu D. Patient Perspectives On Lifestyle Changes Following A Diabetes Diagnosis. KMAN Counseling & Psychology Nexus. 2024;2(1):56-62. [DOI]
4. Tosur M. Inaccurate Diagnosis Of Diabetes Type In Youth: Prevalence, Characteristics, And Implications. Scientific Reports. 2024;14(1). [PMID: 38632329] [PMCID: PMC11024140] [DOI]
5. Padmalal S, Arumugam P, Anusha MB, Bamane KD, Yamsani N, Muppavaram K. Machine Learning For Early Detection Of Chronic Diseases: A Case Study In Diabetes Prediction. Machine Learning in Healthcare: Data-Driven Decisions, Predictive Modelling, Personalized Medicine2026. p. 171. [DOI]
6. Hossain MR, Hossain MJ, Rahman MM, Alam MM. Machine Learning-Based Prediction And Insights Of Diabetes Disease: Pima Indian And Frankfurt Datasets. Journal of Mechanics of Continua and Mathematical Sciences. 2025;20(1):99-114. [DOI]
7. Srivastava S, Sharma L, Sharma V, Kumar A, Darbari H. Prediction Of Diabetes Using Artificial Neural Network Approach. Engineering Vibration, Communication and Information Processing: ICoEVCI 2018, India: Springer Singapore; 2019. p. 679-87. [DOI]
8. Aggarwal Y, Das J, Mazumder PM, Kumar R, Sinha RK. Heart Rate Variability Features From Nonlinear Cardiac Dynamics In Identification Of Diabetes Using Artificial Neural Network And Support Vector Machine. Biocybernetics and Biomedical Engineering. 2020;40(3):1002-9. [DOI]
9. Pei D, Yang T, Zhang C. Estimation Of Diabetes In A High-Risk Adult Chinese Population Using J48 Decision Tree Model. Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy. 2020;13:4621. [PMID: 33273837] [PMCID: PMC7705272] [DOI]
10. Barik S, Mohanty S, Mohanty S, Singh D. Analysis Of Prediction Accuracy Of Diabetes Using Classifier And Hybrid Machine Learning Techniques. Intelligent and Cloud Computing. Singapore: Springer; 2021. p. 399-409. [DOI]
11. Rahman M, Islam D, Mukti RJ, Saha I. A Deep Learning Approach Based On Convolutional Lstm For Detecting Diabetes. Computational Biology and Chemistry. 2020;88:107329. [PMID: 32688009] [DOI]
12. Jaiswal S, Gupta P. Diabetes Prediction Using Bi-Directional Long Short-Term Memory. SN Computer Science. 2023;4(4):373. [DOI]
13. Aslan MF, Sabanci K. A Novel Proposal For Deep Learning-Based Diabetes Prediction: Converting Clinical Data To Image Data. Diagnostics. 2023;13(4):796. [PMID: 36832284] [DOI]
14. Zhang Z, Lu Y, Ye M, Huang W, Jin L, Zhang G, et al. A Novel Evolutionary Ensemble Prediction Model Using Harmony Search And Stacking For Diabetes Diagnosis. Journal of King Saud University—Computer and Information Sciences. 2024;36(1):101873. [DOI]
15. Olabanjo O, Wusu A, Mazzara M. Deep Unsupervised Machine Learning For Early Diabetes Risk Prediction Using Ensemble Feature Selection And Deep Belief Neural Networks. 2023. [DOI]
16. Ahmad HF, Mukhtar H, Alaqail H, Seliaman M, Alhumam A. Investigating Health-Related Features And Their Impact On The Prediction Of Diabetes Using Machine Learning. Applied Sciences. 2021;11(3):1173. [DOI]
17. Awad SF, Dargham SR, Toumi AA, Dumit EM, El-Nahas KG, Al-Hamaq AO, et al. A Diabetes Risk Score For Qatar Utilizing A Novel Mathematical Modeling Approach To Identify Individuals At High Risk For Diabetes. Scientific Reports. 2021;11(1):1-10. [PMID: 33469048] [PMCID: PMC7815783] [DOI]
18. Choubey DK, Kumar P, Tripathi S, Kumar S. Performance Evaluation Of Classification Methods With Pca And Pso For Diabetes. Network Modeling Analysis in Health Informatics and Bioinformatics. 2020;9(1):1-30. [DOI]
19. Sivaranjani S, Ananya S, Aravinth J, Karthika R. Diabetes Prediction Using Machine Learning Algorithms With

Feature Selection And Dimensionality Reduction. 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS): IEEE; 2021. p. 141-6. [DOI]
20. Li X, Zhang J, Safara F. Improving The Accuracy Of Diabetes Diagnosis Applications Through A Hybrid Feature Selection Algorithm. Neural Processing Letters. 2021:1-17. [PMID: 33814965] [PMCID: PMC7997791] [DOI]
21. Gupta S, Sedamkar RR. Genetic Algorithm For Feature Selection And Parameter Optimization To Enhance Learning On Framingham Heart Disease Dataset. Intelligent Computing and Networking. Singapore: Springer; 2021. p. 11-25. [DOI]
22. Chaudhuri A, Sahu TP. Feature Selection Using Binary Crow Search Algorithm With Time Varying Flight Length. Expert Systems with Applications. 2020:114288. [DOI]
23. Khurmaa RA, Aljarah I, Sharieh A. An Intelligent Feature Selection Approach Based On Moth Flame Optimization For Medical Diagnosis. Neural Computing and Applications. 2020:1-40. [DOI]
24. Algamal ZY, Qasim MK, Lee MH, Ali HTM. Improving Grasshopper Optimization Algorithm For Hyperparameters Estimation And Feature Selection In Support Vector Regression. Chemometrics and Intelligent Laboratory Systems. 2021;208:104196. [DOI]
25. Heidari AA, Mirjalili S, Faris H, Aljarah I, Mafarja M, Chen H. Harris Hawks Optimization: Algorithm And Applications. Future Generation Computer Systems. 2019;97:849-72. [PMCID: PMC7296336] [DOI]
26. Elgamal ZM, Yasin NBM, Tubishat M, Alswaitti M, Mirjalili S. An Improved Harris Hawks Optimization Algorithm With Simulated Annealing For Feature Selection In The Medical Field. IEEE Access. 2020;8:186638-52. [DOI]
27. Kaveh A, Eslamlou AD. Water Strider Algorithm: A New Metaheuristic And Applications. Structures. 2020;25:520-41. [DOI]
28. Al-Tawil M, Mahafzah BA, Al Tawil A, Aljarah I. Bio-Inspired Machine Learning Approach To Type 2 Diabetes Detection. Symmetry. 2023;15(3):764. [DOI]
29. Abousaber I. Enhanced Diabetes Prediction Through Advanced Machine Learning And Imbalance Handling Techniques. 2025 4th International Conference on Computing and Information Technology; 2025/04: IEEE; 2025. p. 708-16. [DOI]
30. Shafqat W, Byun YC. A Hybrid Gan-Based Approach To Solve Imbalanced Data Problem In Recommendation Systems. IEEE Access. 2022;10:11036-47. [DOI]
31. Rehman A, Naz S, Razzak I. Leveraging Big Data Analytics In Healthcare Enhancement: Trends, Challenges And Opportunities. Multimedia Systems. 2021:1-33. [DOI]
32. Sarangi AK, Mohapatra AG, Mishra TC, Keswani B. Healthcare 4.0: A Voyage Of Fog Computing With Iot, Cloud Computing, Big Data, And Machine Learning. Fog Computing for Healthcare 40 Environments. Cham: Springer; 2021. p. 177-210. [DOI]
33. Guevara L, Auat Cheein F. The Role Of 5G Technologies: Challenges In SMART Cities And Intelligent Transportation Systems. Sustainability. 2020;12(16):6469. [DOI]
34. Venkatachalam K, Prabu P, Alluhaidan AS, Hubálovský S, Trojovský P. Deep Belief Neural Network For 5G Diabetes Monitoring In Big Data On Edge Iot. Mobile Networks and Applications. 2022:1-10. [DOI]
35. Jha R, Bhattacharjee V, Mustafi A. Classification Of Big Data Using Spark Framework. Proceedings of the Fourth International Conference on Microelectronics, Computing and Communication Systems. Singapore: Springer; 2021. p. 847-54. [DOI]
36. Srivani B, Sandhya N, Rani BP. An Effective Model For Handling The Big Data Streams Based On The Optimization-Enabled Spark Framework. Intelligent System Design. Singapore: Springer; 2021. p. 673-96. [DOI]
37. Suneetha V, Suresh S, Jhananie V. A Novel Framework Using Apache Spark For Privacy Preservation Of Healthcare Big Data. 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA): IEEE; 2020. p. 743-9. [DOI]
38. Ahmed H, Younis EM, Hendawi A, Ali AA. Heart Disease Identification From Patients' Social Posts, Machine Learning Solution On Spark. Future Generation Computer Systems. 2020;111:714-22. [DOI]
39. Tadist K, Mrabti F, Nikolov NS, Zahi A, Najah S. Sdpso: Spark Distributed Pso-Based Approach For Feature Selection And Cancer Disease Prognosis. Journal of Big Data. 2021;8(1):1-22. [DOI]
40. Selvi RT, Muthulakshmi I. Modelling The Map Reduce Based Optimal Gradient Boosted Tree Classification Algorithm For Diabetes Mellitus Diagnosis System. Journal of Ambient Intelligence and Humanized Computing. 2020:1-14. [DOI]
41. Neto C, Senra F, Leite J, Rei N, Rodrigues R, Ferreira D, et al. Different Scenarios For The Prediction Of Hospital Readmission Of Diabetic Patients. Journal of Medical Systems. 2021;45(1):1-9. [PMID: 33409739] [DOI]
42. Kour H, Sabharwal M, Suvanov S, Anand D. An Assessment Of Type-2 Diabetes Risk Prediction Using Machine Learning Techniques. Proceedings of International Conference on Big Data, Machine Learning and their Applications. Singapore: Springer; 2021. p. 113-22. [DOI]
43. Al-Hameli BA, Alsewari AA, Alsarem M. Prediction Of Diabetes Using Hidden Naïve Bayes: Comparative Study. Advances on Smart and Soft Computing. Singapore: Springer; 2021. p. 223-33. [DOI]
44. Kurt B, Gürlek B, Keskin S, Özdemir S, Karadeniz Ö, Kırkbir IB, et al. Prediction Of Gestational Diabetes Using Deep Learning And Bayesian Optimization And Traditional Machine Learning Techniques. Medical & Biological Engineering & Computing. 2023:1-12. [PMID: 36848010] [PMCID: PMC9969040] [DOI]
45. Tasin I, Nabil TU, Islam S, Khan R. Diabetes Prediction Using Machine Learning And Explainable Ai Techniques. Healthcare Technology Letters. 2023;10(1-2):1-10. [PMID: 37077883] [PMCID: PMC10107388] [DOI]
46. Vinoth N, Vijayakarthick M, Ramesh S, Sivaraman E. Taxonomy Of Diabetic Retinopathy Patients Using Biogeography-Based Optimization On Support Vector Machine Based On Digital Retinal Images. Sustainable Communication Networks and Application. Singapore: Springer; 2021. p. 631-42. [DOI]
47. Akbar S. A Hybrid Ensemble Feature Selection-Based Segmentation And Deep Majority Voting Framework On Large Multi-Class Diabetes Retinopathy Databases. Turkish Journal of Computer and Mathematics Education (TURCOMAT). 2021;12(12):416-28.
48. Luo X, Wang W, Xu Y, Lai Z, Jin X, Zhang B, et al. A Deep Convolutional Neural Network For Diabetic Retinopathy Detection Via Mining Local And Long-Range Dependence. CAAI Transactions on Intelligence Technology. 2023. [DOI]
49. Altobelli E, Angeletti PM, Marziliano C, Mastrodomenico M, Giuliani AR, Petrocelli R. Potential Therapeutic Effects Of Curcumin On Glycemic And Lipid Profile In Uncomplicated Type 2 Diabetes-A Meta-Analysis Of Randomized Controlled Trial. Nutrients. 2021;13(2):404. [PMID: 33514002] [PMCID: PMC7912109] [DOI]
50. Zhang Z, Ahmed KA, Hasan MR, Gedeon T, Hossain MZ. A Deep Learning Approach To Diabetes Diagnosis. Asian Conference on Intelligent Information and Database Systems; 2024/04: Springer Nature Singapore; 2024. p. 87-99. [DOI]


Health Nexus
E-ISSN: 2981-2569

51. Khaleel FA, Al-Bakry AM. Diagnosis Of Diabetes Using Machine Learning Algorithms. Materials Today: Proceedings. 2023;80:3200-3. [DOI]



Health Nexus
E-ISSN: 2981-2569