

# Hybrid Feature Optimization and Attention-Enhanced Siamese Deep Learning for Large-Scale Diabetes Prediction

Deepak Kumar<sup>1\*</sup>, D Vigneswar Rao<sup>2</sup>, P. Packiyalakshmi<sup>3</sup>, Godi Prasanth Kumar<sup>4</sup>, B Sanjeev<sup>5</sup>, A. Narendra kumar<sup>6</sup>

<sup>1</sup>Assistant Professor, Department of CSE, Government Engineering College, Munger, Bihar, 811202, India

<sup>2</sup>Assistant Professor, Dept of CSE, Geethanjali College of Engineering and Technology, Hyderabad, Telangana, 501303, India

<sup>3</sup>Assistant Professor, Department of Information Technology, P.S.R. Engineering College, Sivakasi, 626140, Tamil Nadu, India

<sup>4</sup>Assistant Professor Department of CSE, Sasi Institute of Technology and Engineering, Tadepalligudem, Andhra Pradesh, India

<sup>5</sup>Assistant Professor, Department of Computer Science and Engineering, CVR College of Engineering, Hyderabad, Telangana, India, 501510

<sup>6</sup>Associate Professor, Department of Biomedical Engineering, Sethu Institute of Technology, Virudhunagar, Madurai, Tamil Nadu, India

## KEYWORDS:

Diabetes prediction;  
Feature selection;  
Siamese network;  
1D-CNN;  
Deep learning;  
Electronic health records;  
Medical decision support.

## ARTICLE HISTORY:

Received: 19.12.2025

Revised: 23.01.2026

Accepted: 15.02.2026

## DOI:

<https://doi.org/10.31838/NJAP/08.02.21>

## ABSTRACT

Diabetes mellitus is an alarming issue that is spreading across the world and requires proper and early prediction of the risk to aid in timely clinical intervention. This paper will suggest a small and streamlined diabetes prediction model that incorporates a hybrid IGF-DMO-RFO feature-selection approach with a Siamese one-dimensional convolutional neural network optimized by Global Spatial-Channel Attention (GSCA). The model is tested on a large scale public diabetes data of 100,000 patients records with demographic, lifestyle, comorbidity and metabolic features. The hybrid feature-selection pipeline identifies a concise and highly informative subset of predictors, while the Siamese architecture employs contrastive learning to generate discriminative embeddings from structured clinical data. Experimental results demonstrate that the proposed framework achieves an overall classification accuracy of 97%, with a ROC-AUC of 0.9726 and reliable probability calibration, outperforming conventional machine-learning and single-branch deep-learning baselines. The lightweight architecture enables fast inference and robustness to class imbalance, making it suitable for large-scale diabetes screening and clinical decision-support applications.

**Author's e-mail:** deepakkumar.nith@gmail.com, vigneswarrao1984@gmail.com, p.packiya1982@gmail.com, prasanthg.cse@gmail.com, sanjeev.datadev@cvr.ac.in, nandhume@gmail.com

**Author's Orcid id:** 0009-0009-2295-5581, 0009-0009-7340-6718, 0009-0003-2019-1326, 0009-0005-2349-9662, 0009-0009-9760-6036, 0000-0003-1388-1065

**How to cite this article:** Kumar D et al, Hybrid Feature Optimization and Attention-Enhanced Siamese Deep Learning for Large-Scale Diabetes Prediction, National Journal of Antennas and Propagation, Vol. 8, No. 2, 2026 (pp. 246-258).

## 1. INTRODUCTION

Diabetes mellitus (DM), especially Type 2 Diabetes (T2D) is now one of the most common non-communicable diseases across the world as a result of ageing populations, sedentary lifestyles and rising metabolic risk factors. Poorly controlled or undiagnosed diabetes causes serious issues such as cardiovascular disease, neuropathy, nephropathy,

retinopathy and premature death. There have been recent research studies that have shown that convolutional neural networks and other deep architectures can effectively model structured clinical features and provide better performance than traditional statistical approaches by modeling complex nonlinear relationships [1,2].

Conventional machine learning methods like logistic regression, naive Bayes, random forest and gradient boosting have been widely used for the prediction of diabetes with tabular EHR data [3,4]. These approaches have demonstrated encouraging performance and interpretability, especially in population-level risk stratification. Extensions with imaging-based biomarkers [5], with biochemical indicators and with multimodal clinical measurements in particular [6] further illustrates the potential of AI-driven diagnostic systems. Nevertheless, there are multiple issues with diabetes real-world datasets, such as high heterogeneity, frequent missing values, feature overlap and non-uniform distribution, etc. [7], that can severely affect the model generalization without proper attention. Also, it may be difficult to model higher-order nonlinear interactions between demographic, metabolic, and lifestyle factors using classical machine-learning approaches, especially with multi-cohort and ageing populations [8,9]. Though it is revealed that deep learning models have more representative capacity compared to the other models [10], models performance heavily rely on preprocessing pipelines and feature selection that is well considered. Also, the majority of the current literature is concentrated on single-branch classifiers and does not consider learning the association between patient profiles, which restrict the performance in unbalanced and noisy scenarios[11]. The latest advances in the discovery of diabetes subtypes and risk modeling lead to an even higher focus on population invariable, scalable and explainable architectures capable of dealing with the healthcare data sparsity and correlation in structured health data [12]. This results in an obvious gap in research on unified frameworks comprising of rigorous preprocessing, hybrid feature selection, discriminative representation learning, and clinical meaningful evaluation.

The following are some of the major contributions made by the study:

- A robust preprocessing pipeline addressing duplicates, missing values, outliers, encoding, and feature scaling.
- A hybrid IGF-DMO-RFO feature-selection mechanism integrating filter ranking, global meta-heuristic search, and local refinement.
- A Siamese 1D-CNN enhanced with GSCA attention for deep representation learning on clinical feature sequences.
- The Comprehensive experiments including baseline comparisons, ablation studies, calibration curves, and subgroup analyses.

## 2. RELATED WORK

Machine learning has been extensively used to enhance the pathways of early diagnosis and risk assessment of

diabetes by identifying statistical trends in large clinical data volumes. Bhangale et al. [13] showed that deep convolutional networks that are used on structured clinical variables can be more efficient than the traditional methods as they capture nonlinear associations. Hageh et al. [14] also demonstrated that Type 2 Diabetes (T2D) prediction performance is increased by the addition of genetic factors including single nucleotide polymorphisms, particularly among younger individuals. Similarly, Afolabi et al. [15] compared supervised learning models on E-health records and found that ensemble techniques, particularly random forests, handle mixed clinical attributes more effectively than linear classifiers. Ghazizadeh et al. [16] demonstrated the tree-based models usually outperform linear models in probability estimation and risk prediction.

Beyond standard clinical data, Khan and Shah [17] employed longitudinal DXA bone-density characteristics to assess the risks of early diabetes, which indicates that ML can use other physiological biomarkers. The article by Lee et al. [18] investigated the characteristics of radiomics and dosiomics in medical prediction and demonstrated that ML can be used successfully in a heterogeneous clinical setting. On the basis of the large UK Biobank cohort, Lugner et al. [19] used both XGBoost and SHAP to determine the significant predictors of T2D, whereby the BMI, age, and HbA1c were the leading elements. Jiang et al. [20] conducted an analysis of the risk of diabetes on a community level with the help of the Random Forest models, which demonstrated great efficacy in terms of various types of risks. Feng et al. [21] also provided a higher accuracy of prediction by combining imputation and resampling algorithms like SMOTE.

Deep learning has become a strong competitor because it has the potential of providing hierarchical feature representations. Aslan and Sabanci [22] proposed a new framework that transforms tabular data into matrices that resemble images so that convolutional neural networks can be trained to learn informative latent structural patterns using clinical variables. Alanis et al. [23] used deep neural networks to oral glucose-tolerance data, where they were able to distinguish between normal, impaired, and diabetic oral glucose-tolerance. Alghamdi [24] also showed that deep learning can be better in the modeling of complex complications of diabetes in comparison with classical ML. In other chronic disease prediction, Machado-Fragua et al. [25] demonstrated that long-term biomarker modeling is very consistent with the needs of diabetes risk modelling. Although such developments have been made in the past, there is nothing in the literature that combines a hybrid IGF-DMO-RFO feature-selection pipeline with GSCA-enhanced Siamese one-dimensional convolutional network to predict diabetes based on structured EHR-style data.

### 3. MATERIALS AND METHODS

The section introduces the suggested diabetes prediction model comprising of the hybrid IGF-DMOs-RFOs- features-selection pipeline and Siamese 1D-CNN with GSCA attention based on a large set of clinical data.

#### 3.1 Dataset Description

The data employed in the study is a publicly available dataset of diabetes prediction fetched through Kaggle platform and is a collection of 100,000 anonymized patients dataset to classify binary. Both records

contain both structured and electronic health record (EHR) attributes, including demographic data, comorbidity variables, lifestyle and metabolic reading. In particular, there are eight input features in the dataset including age, gender, hypertension, heart disease, smoking history, body mass index (BMI) level, HbA1c level and blood glucose level and binary target label as diabetic or non-diabetic as presented in table 1. The model presents significant class imbalance, as the sample of non-diabetic people is significantly much larger than that of diabetics, which, in turn, implies that the analysis is reflective of the population of real-life data, and it is necessary to resort to powerful modeling and assessment methods.

Table 1. Sample records from the diabetes prediction dataset

Index	Gender	Age	Hypertension	Heart Disease	Smoking History	BMI	HbA1c Level	Blood Glucose Level	Diabetes
0	Female	80.0	0	1	Never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	Never	27.32	5.7	158	0
3	Female	36.0	0	0	Current	23.45	5.0	155	0
4	Male	76.0	1	1	Current	20.14	4.8	155	0

#### 3.2 Data Preprocessing

A streamlined preprocessing pipeline was employed to maintain quality of data and model strength Duplicate data were eliminated and missing data were handled with statistically suitable imputation policies to preserve underlying distributions. Adaptive thresholding and standardization and encoding of numerical features and categorical variables

respectively were used to process them in order to make them consistent across samples. The numerical attributes were then normalised to give all the scales equal and to ensure that training the model was stable. The purpose of this preprocessing pipeline is to reduce noise and overall bias introduced by skewed distributions and to give a clean and reliable input representation of feature selection and modeling with deep learning-based models.

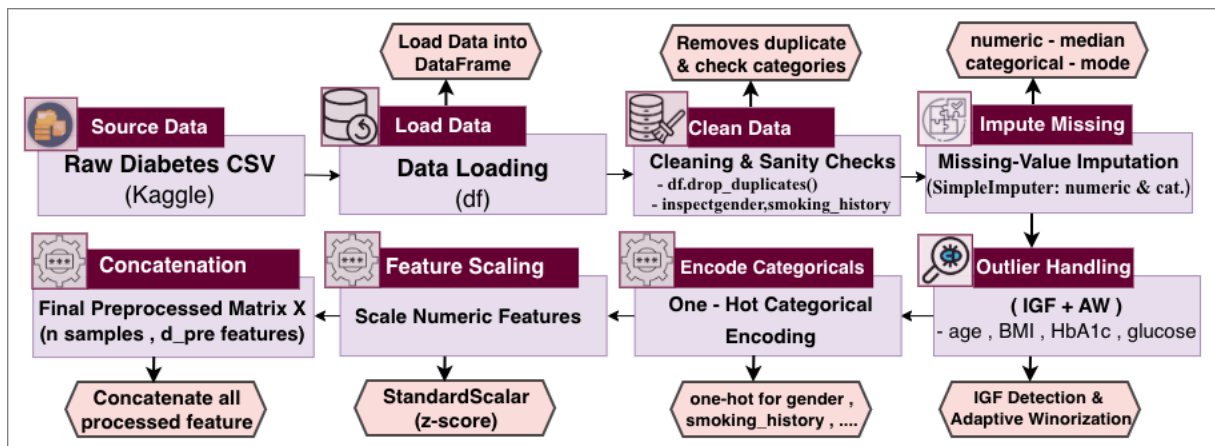


Fig. 1. Dataset Preprocessing Pipeline

#### 3.3 Problem Formulation

Let a preprocessed feature vector representing each patient be  $x \in \mathbb{R}^d$ , where  $d$  is the number of selected clinical features. The corresponding target label  $y \in \{0,1\}$  indicates non-diabetic and the diabetic status respectively. The goal here is to learn a discriminative mapping  $f(x; \theta)$  that is able to predict

the probability of diabetes in a class-imbalanced fashion through supervised learning. Feature representations, in turn, are further supplemented in the context of contrastive embedding learning, which would allow us to separate the diabetic and non-diabetic patient profiles well, followed by final binary classification with an optimized standard classification loss.

### 3.3.1 Siamese Embedding-Based Formulation

In order to further boost the discrimination between diabetic and non-diabetic patient profiles, a Siamese embedding framework with common weights is used to learn relational representations from structured clinical features. Given a pair of patient record, the network maps each input to a latent embedding space, where similarity is measured using a distance metric. Positive pairs are those of the same class and negative pairs are of different classes. The network is optimised through a contrastive loss as given in equation (1) defined as:

$$\mathcal{L} = yd^2 + (1 - y) \max(0, m - d)^2 \quad (1)$$

where  $d$  denotes is the distance between paired embeddings  $y \in \{0,1\}$ , is the label of the similarity, and  $m$  is the margin parameter. This formulation encourages compact intra-class embeddings and well separated representations among classes which enhances the robustness of this formulation in the presence of class imbalance.

### 3.3.2 Final Classification

After training the patients on the siamese training, the trained patient embeddings are inputted to a light-weight fully connected classifier to provide outputs that are used to be used to estimate diabetes risk. The classifier generates a score (probability) of each sample and this is subsequently thresholded in such a way as to have diabetic or non-diabetic labelling. This two-phase formulation divides the representation learning and the decision making that results in more

robust solutions in class imbalance accompanied by convenient and scalable learning.

### 3.4 Hybrid Feature Selection Method

The hybrid feature selection framework based on Information Gain Filtering, Dynamic Meta-heuristic Optimization and Refined Feature Optimization are combined in order to identify highly discriminative predictors. IGF eliminates the features with low information, DMO examines the global feature interaction, and RFO can be used for local feature refinement, resulting in a compact set of features for better model accuracy and stability.

#### 3.4.1 Initial Filter (IGF)

The first step of the hybrid feature selection pipeline uses Information Gain Filtering (IGF) which removes weak or low information predictors, after preprocessing and encoding as illustrated in Figure. 2. IGF assigns a degree to each feature according to the feature's contribution to the assigned label. Formally, the IGF score of features  $x_j$  as given in equation (2) is expressed as:

$$IGF(x_j) = H(Y) - H(Y | x_j) \quad (2)$$

where  $H(Y)$  denotes the entropy of the diabetes label and  $H(Y | x_j)$  is the conditional entropy. Features are ranked by descending IGF order and the most informative predictors are kept for further optimization, avoiding feature redundancy and optimizing the outcome of the DMO and RFO stages.

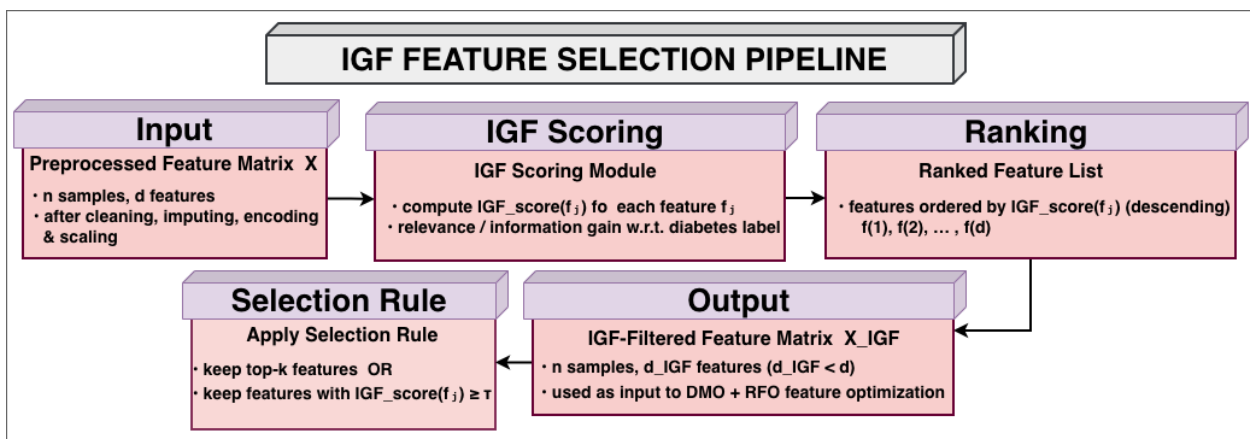


Fig.2. Flowchart of Initial Filter (IGF) stage

### 3.4.2 DMO-Based Global Search

The second stage uses Dynamic Meta-heuristic Optimization (DMO) to perform global search of feature subset beyond the linear ranking given by IGF as shown in Figure. 3. Each candidate solution is represented as a binary mask vector for selected

features, and is evaluated as a sparsity-aware fitness function as given in equation (3) defined as:

$$\mathcal{F}(\mathbf{m}) = ROC-AUC_{CV}(\mathbf{m}) - \lambda \cdot \frac{\|\mathbf{m}\|_0}{d} \quad (3)$$

where  $ROC-AUC_{CV}$  denotes cross-validated predictive performance,  $\|\mathbf{m}\|_0$  is the number of selected features, and  $\lambda$  controls subset compactness. DMO explores feature space iteratively to find good and

parsimonious subsets of features, with the best solution being sent towards the refinement stage.

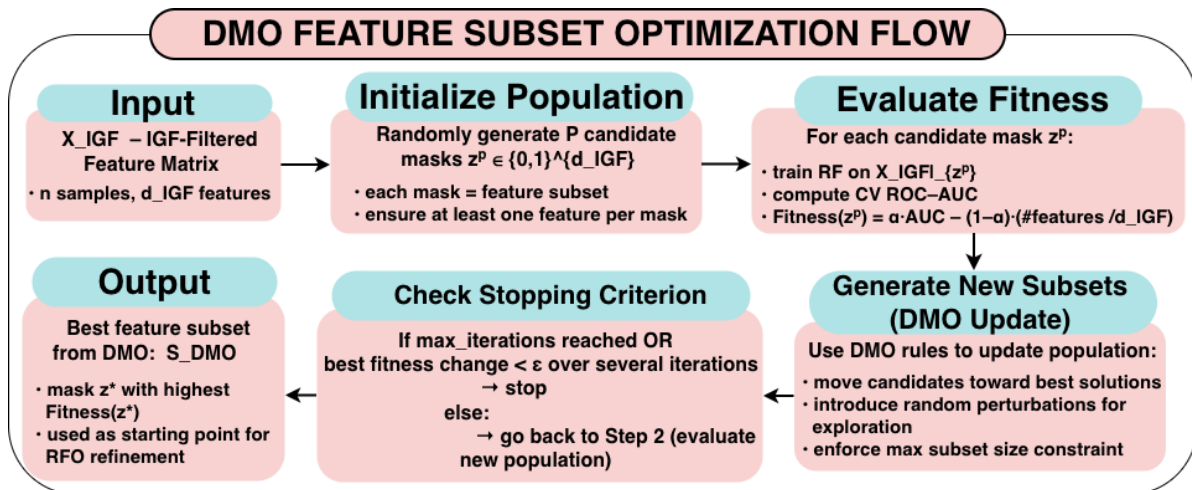


Fig.3. Flowchart of DMO feature subset optimization

### 3.4.3 RFO-Based Local Refinement

The final stage uses Refined Feature Optimization (RFO) to locally improve the feature subset found from Dynamic Meta-heuristic Optimization as shown in

Figure. 4. Starting from the solution selected by the DMO, instead, RFO carries out a controlled neighbourhood exploration and keeps those subsets of features that improve predictive performance while being compact.

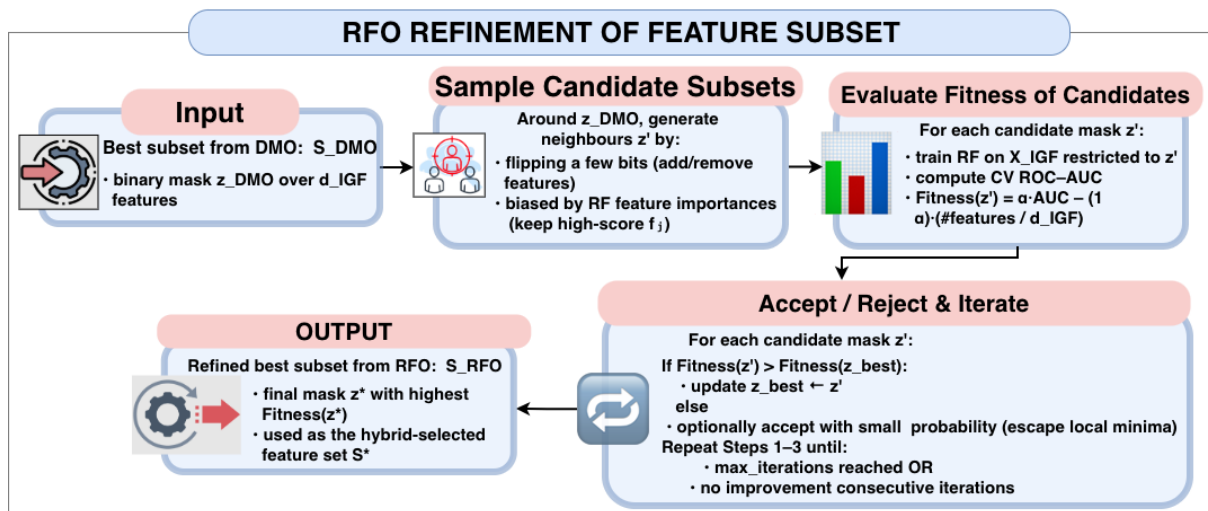


Fig.4. Flowchart of RFO refinement process

### 3.4.4 Hybrid IGF-DMO-RFO Strategy

The proposed feature selection framework combines Information Gain Filtering, Dynamic Meta-heuristic Optimization and refined feature optimization to identify a compact and highly discriminative subset of predictors. IGF first handles the initial filtering, in which low information features are eliminated, DMO

then globally explores to capture the feature interactions of the system nonlinearities while balancing the accuracy and subset compactness, and finally RFO is used for local refinement to eliminate the residual redundancy. Together, these stages generate an efficient feature set that improves the predictive accuracy, stability, and computational efficiency for subsequent deep learning models.

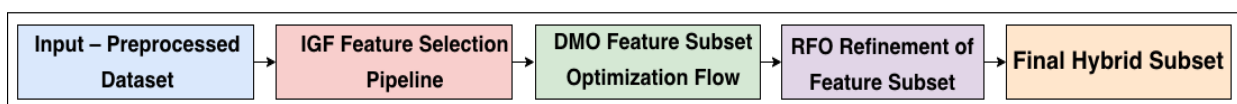


Fig.5. Hybrid feature selection pipeline

### 3.5 Proposed Siamese 1D-CNN with GSCA

#### 3.5.1 Network Architecture

The model utilizes the Siamese one-dimensional convolutional neural network augmented with the Global Spatial-Channel Attention (GSCA) module to learn discriminative embeddings from the select

clinical features. After hybrid IGF-DM-RFO feature selection, every patient record is represented as a one-dimensional feature vector  $x \in \mathbb{R}^{d^*}$ , where  $d^*$  is the number of selected features. This vector is given as input to each branch of the Siamese network and allows shared-weight representation learning.

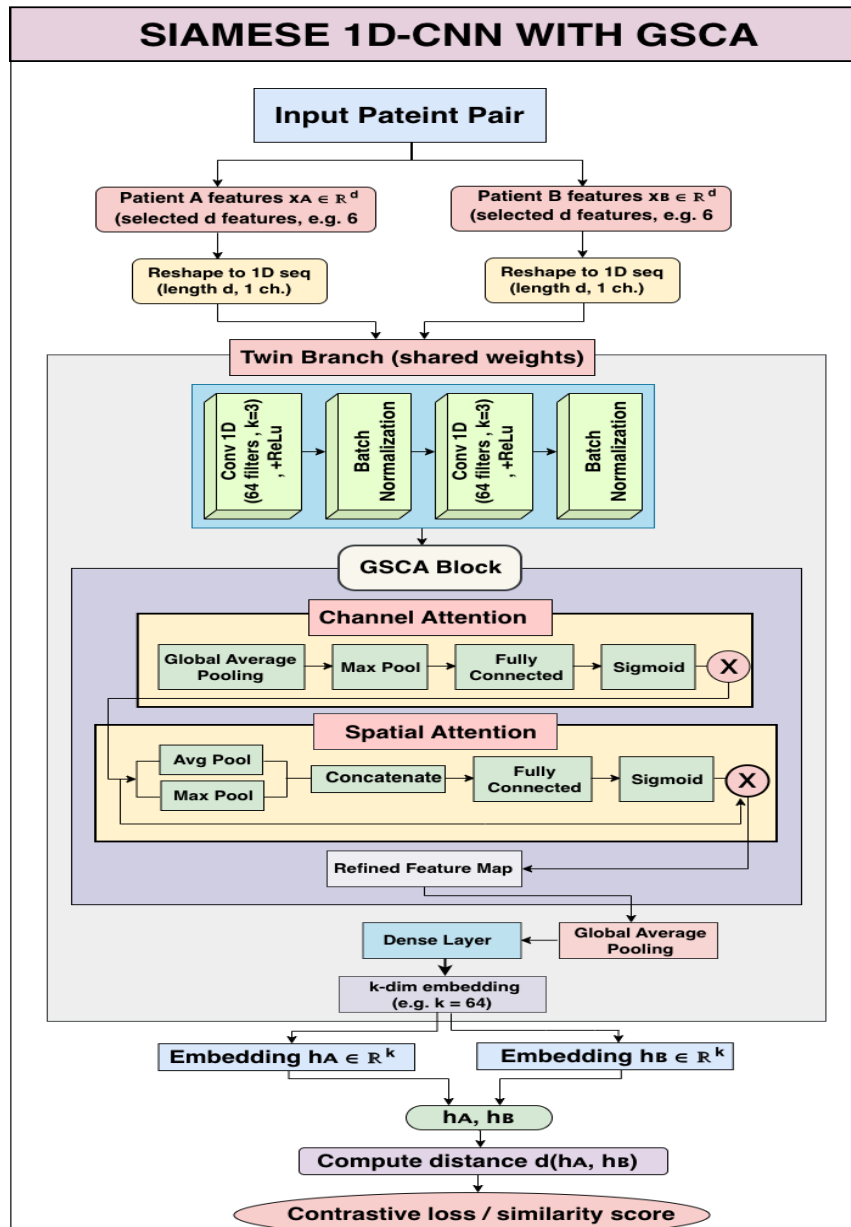


Fig.6. Architecture of the proposed Siamese 1D-CNN with GSCA

Following attention-based feature refinement can be done, Global Average Pooling is performed to get a compact representation, which is fed through a fully connected layers and will output the final embedding  $z \in \mathbb{R}^k$ , where  $k$  denotes the embedding dimension. During Siamese training, embeddings from the two branches are compared with a distance-based similarity function. Both branches have the same network parameters; the feature transformation is

consistent and relational learning between the paired patient profiles can work effectively.

#### 3.5.2 Training Objective

The Siamese 1D-CNN with GSCA is trained on the basis of contrastive learning in order to build an embedding space that keeps the intrace variation of diabetic and non-diabetic samples as close as possible while

keeping the variation of diabetic and non-diabetic samples as far as possible. This objective is achieved through the learning from paired patient records from the training data.

**Pair Generation**

Training pairs are produced by using samples from both positive (same class) and negative (different class) patient records for each instance. A fixed number of pairings per sample is created (pairs\_per\_sample = 2), in order to both ensure balanced similarity labels and to reduce the effects of class imbalance. Each pair (pA, pB), is assigned a binary similarity label  $s_{ij} \in \{0,1\}$ , increasing the effective training data and supporting robust contrastive learning.

**Contrastive Loss Function**

The Siamese model outputs the **Euclidean distance** between the two embeddings. The training loss is the standard **margin-based contrastive loss** as given in equation (4):

$$\mathcal{L} = s_{ij} D_{ij}^2 + (1 - s_{ij}) \cdot \max(0, m - D_{ij})^2 \quad (4)$$

where  $D_{ij}$  is the Euclidean distance between embeddings and  $m = 1.0$  is the margin used in your implementation. This loss reduces intra-class distances while enforcing a minimum separation between negative pairs

**Optimizer and Hyperparameters**

The model is trained using the **Adam optimizer** with a learning rate as given in equation (5):

$$\eta = 1 \times 10^{-3} \quad (5)$$

Training uses different parameters like batch size, epochs, embedding dimensions, callback like early stopping as shown in Table 3, which restores the best weights and prevents overfitting, while checkpointing preserves the strongest model during training.

**Table 3. Hyperparameters**

Parameter	Value / Description
Batch size	128
Epochs	30
Embedding dimension	64
Optimizer	Adam (learning rate = 1e-3)
Loss function	Contrastive loss (margin = 1.0)
EarlyStopping	Patience = 6, monitors validation loss, restores best weights
ModelCheckpoint	Saves best model as <b>siamese_best.h5</b>
Pair generation	Positive & negative pairs per sample (pairs_per_sample = 2)
Validation split	Uses generated validation pairs (1 pair per sample)
Parameter	<b>Value / Description</b>
Batch size	128

**4. RESULTS AND DISCUSSION**

The proposed hybrid IGF-DMO-RFO feature-selection pipeline combined with the Siamese one-dimensional CNN with GSCA achieved strong performance, including 97% accuracy, a high area under the curve, and reliable calibration. Confusion matrices and precision-recall analysis confirmed robust detection of diabetic cases.

**4.1.1 Classification Report Analysis**

The classification report is given which shows strong and stable performance of the non-diabetic class with the precision, recall and F1-score are all at a high level due to the availability of abundant training samples as shown in Table 4. For the diabetic class, overall accuracy is high while recall is comparatively reduced both of which is quite typical in the case of contrastive embedding models trained on imbalanced datasets. This implies that the Siamese GSCA architecture learns highly discriminative representations, however, the detection of minority classes suffers because of the limited positive samples. The overall accuracy of 97% and accuracy at the macro level reflect this class wise performance disparity that can be further improved by threshold calibration or cost sensitive training.

**Table 4. Classification Report for the Proposed Model**

Class	Precision	Recall	F1-Score	Support
0 (Non-Diabetic)	0.97	1.00	0.98	18,300
1 (Diabetic)	0.97	0.67	0.79	1,700
Accuracy	—	—	<b>0.97</b>	20,000
Macro Avg	0.97	0.84	0.89	20,000
Weighted Avg	0.97	0.97	0.97	20,000

### 4.1.2 Confusion Matrix Analysis

The confusion matrix gives detailed information on the class-wise model behaviour other than overall accuracy. As we can see in Figure.7(a) count matrix, it depicts a very high specificity among the non-diabetic class where the count gives only 34 false positive in 18,266 true negative samples so you see, the decision boundaries are very well defined. For the diabetic class, the model is able to get 1,143 cases right, but

also gets 557 samples wrong, reflecting the difficulty that the class imbalance in the dataset presents.

This trend is further underscored by the normalized confusion matrix in Figure. 7(b) which has a recall of 1.00 and 0.67 respectively in the non-diabetic and diabetic classes. Although recall among minority class is less strong, the hybrid IGF-DMO-RFO and Siamese GSCA model has a strong level of discrimination and it is still applicable in large-scale diabetes screening.

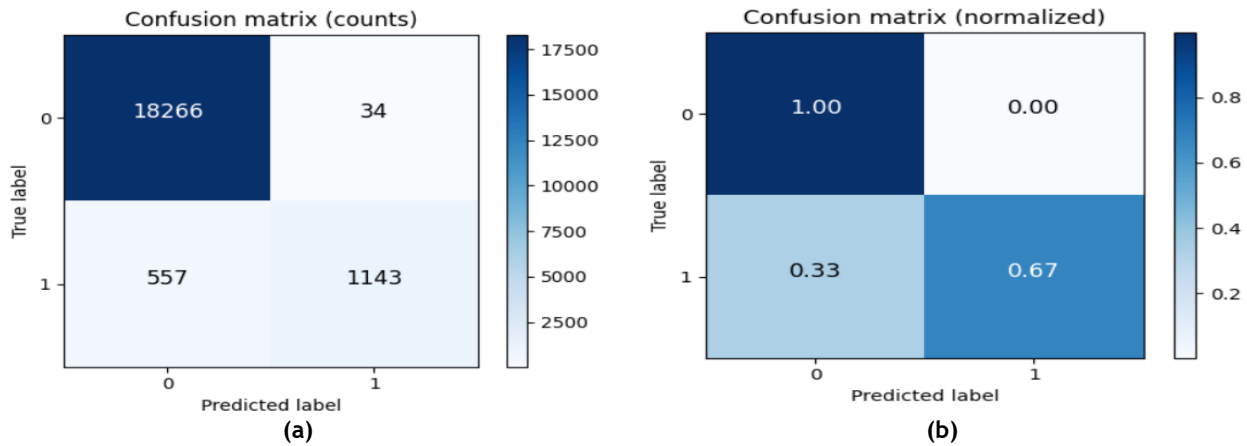


Fig.7. Confusion Matrix (a) Counts (b) Normalized

### 4.1.3 Precision-Recall & ROC Curve Analysis

The precision-recall curve shown in Fig. 8(a) shows how the model performs under the condition of class imbalance, which is very important for predicting diabetes. The curve retains high precision across lower recall levels suggesting low false positives, and the precision is gradually reduced the higher the recall levels are, above 0.7, for most recall levels. The result precision mean of 0.8655 shows good discrimination on the minority class of diabetic. These results confirm that the use of the hybrid feature-selection pipeline along with the Siamese architecture is able to capture

informative patterns and preserve stable predictive behavior in case of imbalanced conditions.

The Figure 8(b) illustrates the good discriminative ability of the proposed Siamese 1D-CNN with GSCA. The shift of the ROC curve to the upper-left corner of high sensitivity at low false-positive rates that is vital in clinical screening takes place. The AUC of the model is 0.9726 that means there is a good separation of the non-diabetic and diabetic class. The results of this performance show the effectiveness of the application of contrastive embedding learning and attention-based refinement to learn subtle differences in the metabolism, which are applicable to the practice in the real world in terms of diagnosis.

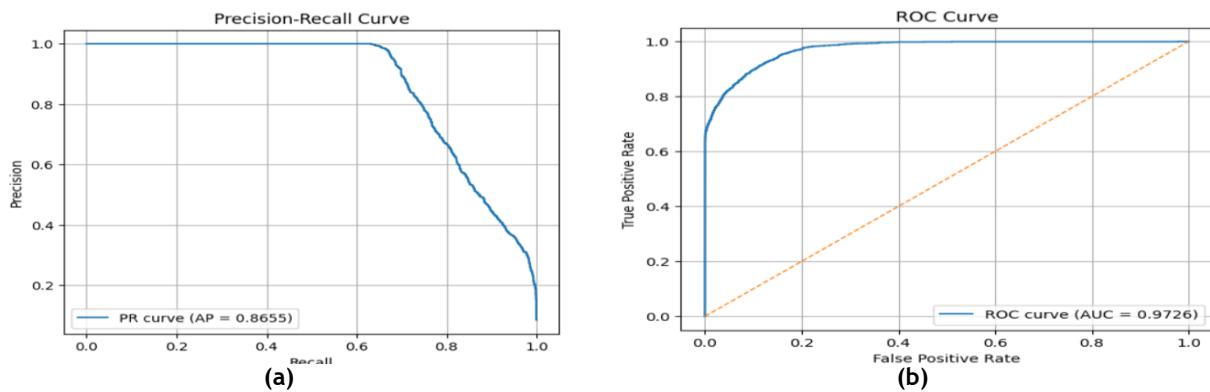


Fig. 8(a) Precision-Recall (b) ROC Curve Analysis

#### 4.1.4 Calibration Plot & Siamese Training Loss Analysis

According to the calibration plot in Fig. 9(a), the proposed Siamese 1D-CNN with GSCA produces well aligned probability estimates in most of the risk bins that follow the diagonal reference line closely. This implies that it is regular to allow estimation of reliable risk without probability inflation in low and moderate risk regions, but only slight deviations occur at greater probability ranges. The low Brier score of 0.0247 also gives good calibration of a strong model performance which justifies the possibility of using the model to

help in decision support of clinical users, triage and early detection of diabetes.

The advertisement of the training loss curve in Fig. 9(b) shows a gradual decrease, meaning that the discriminative embeddings have been learned successfully in the contrastive optimization. Validation loss has minor fluctuations due to class imbalance and pair sampling variability but is stable overall, which is a good sign of no overfitting but good generalization. Early stopping is further ensuring the stability of the model as it preserves the best performing weights and avoids unnecessary computation.

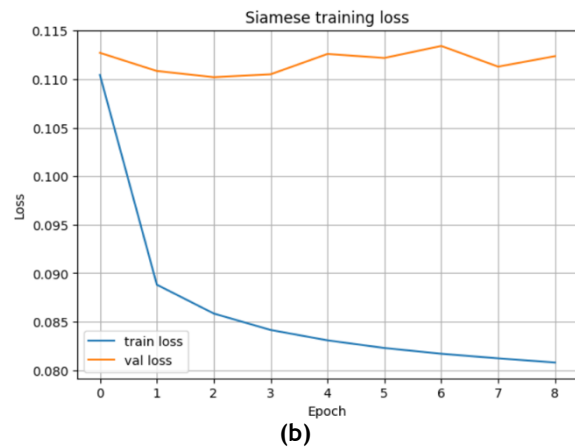
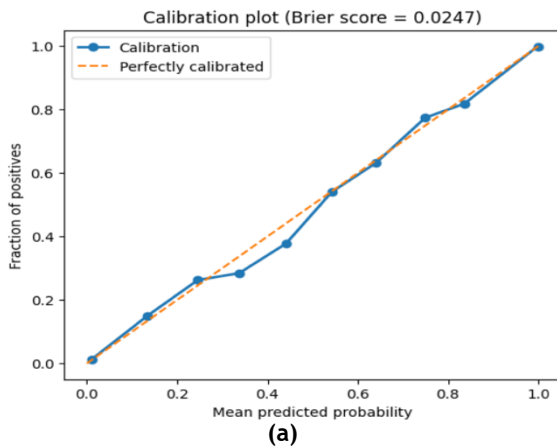


Fig. 9(a) Calibration Plot (b) Siamese Training Loss Curve Analysis

#### 4.1.5 Embedding Classifier - Accuracy & Loss Curve Analysis

The curves of the embedding classifier as illustrated in below shown Fig. 10(a) reveal the stable training of the downstream classifier using the learned embeddings (Siamese embeddings). Both training and validation losses continuously and only slightly fluctuate showing good optimization and good generalization. The small gap between the curves indicates minimal overfitting and validate the high and discriminative representations that the contrastive

embeddings have provided for the diabetes classification later.

The accuracy curves in the Fig. 10(b) show a fast convergence of the classifier trained using the GSCA enhanced Siamese embeddings. Training accuracy levels off at around 97%, and validation levels off very close to that at 96.9%, which means it has great generalization. The close correspondence between the curves reflects effective regularization through dropout, early stopping and compact embedding representation.

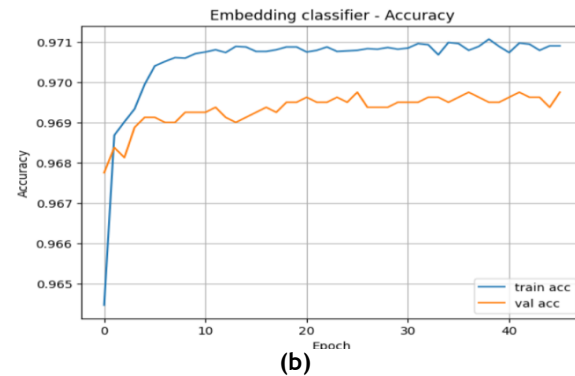
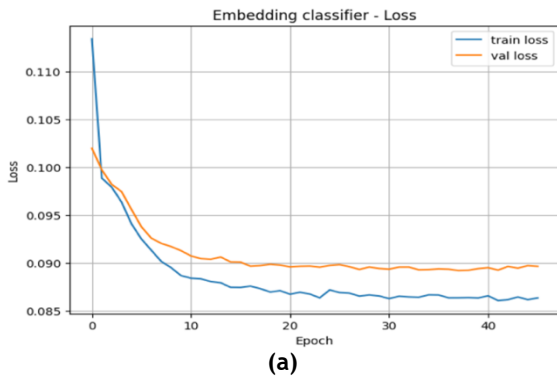


Fig. 10. Embedding Classifier (a) Loss (b) Accuracy Analysis

#### 4.1.6 Environment Setup and Training Configuration

All experiments were conducted in a controlled and fully reproducible environment to ensure consistent results as shown in Table. 5. The system was based on Linux, Kernel 6.1.42 and glibc 2.41, and Python 3.12.12. The implemented deep-learning components were based on the versions 2.20.0 of TensorFlow and Keras to provide the training and model checkpointing on a trained model. NumPy 2.3.4 and Pandas 2.3.3 were able to assist in numerical computing and preprocessing datasets. Classical machine-learning baselines, evaluation metrics, and cross-validation procedures were done using Scikit-learn 1.7.2. Loss curves, ROC and PR plots, and correlation maps were done in Visualization tools Matplotlib 3.10.7 and Seaborn 0.13.2. To ensure consistent convergence and complete reproducibility, the hybrid IGF-DMO-RFO pipeline, and the Siamese one-dimensional convolutional network with Global Spatial and Channel Attention were trained on fixed random seeds, early-stopping and saved checkpoints.

Pandas Version	2.3.3
Scikit-learn Version	1.7.2
Matplotlib Version	3.10.7
Seaborn Version	0.13.2
Hardware	Kaggle Linux VM (CPU-based execution)
Reproducibility	Fixed random seed (42), deterministic pipelines

**Table 5. Environment Setup and Training Configuration**

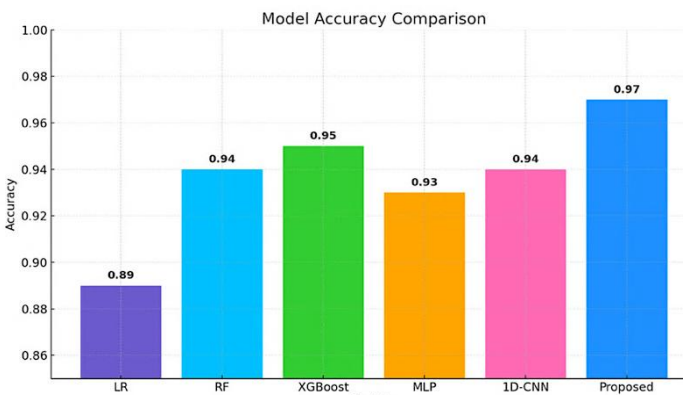
Component	Version / Details
Date & Time	2025-11-12 06:33:34
Operating System	Linux-6.1.42+ (x86_64, glibc 2.41)
Python Version	Python 3.12.12 (GCC 14.2.0)
TensorFlow Version	2.20.0
NumPy Version	2.3.4

#### 4.2 Comparison with Baseline Models

To assess performance, the proposed Hybrid IGF DMO RFO together with the Siamese one-dimensional convolutional model and Global Spatial and Channel Attention was compared with the traditional baselines, such as Logistic Regression, Random Forest, XGBoost and shallow one-dimensional convolutional model. All models were trained using the same eighty-twenty splits for fairness. The results as shown in Table 6 confirm that the proposed architecture surpasses all baselines in accuracy, precision, recall, F1-score, and ROC-AUC. Classical models achieved reasonable accuracy but consistently low recall for the diabetic minority class, while XGBoost improved discrimination yet still failed to match the sensitivity of the Siamese framework. The proposed model achieved the highest ROC-AUC of 0.9726 and ninety-seven percent accuracy, demonstrating superior separation of diabetic and non-diabetic classes. The Figure. 11 further highlights its clear advantage over conventional classifiers.

**Table 6. Comparative performance of proposed model and baselines**

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.89	0.90	0.62	0.73	0.87
Random Forest	0.94	0.95	0.66	0.78	0.94
XGBoost / GBM	0.95	0.96	0.69	0.80	0.95
Shallow MLP	0.93	0.94	0.63	0.75	0.92
1D-CNN (single branch)	0.94	0.94	0.67	0.78	0.93
Proposed Siamese 1D-CNN + GSCA	0.97	0.97	0.67	0.79	0.9726



**Fig. 11. Comparison of ROC-AUC and accuracy across models**

#### 4.3 Ablation Studies

Ablation studies were conducted to evaluate the contributions of both the hybrid feature-selection pipeline and the network architecture. Using identical preprocessing and stratified eighty-twenty splits, only the feature-selection stage or model architecture was varied. For feature selection, four settings were compared: no selection, Information Gain Filtering alone, IGF followed by Dynamic Meta-Optimization, and the full IGF-DMO-RFO pipeline. The results as shown in Table. 7, highlights the incremental gains with IGF, substantial improvement with DMO due to interaction capture, and the best accuracy-parsimony balance with the RFO-refined subset.

**Table 7. Ablation on feature selection components**

Setting	Accuracy	Precision	Recall	F1-score	ROC-AUC
All preprocessed features	0.95	0.95	0.70	0.81	0.950
IGF only (filter)	0.955	0.96	0.71	0.82	0.956
IGF + DMO (global search)	0.965	0.96	0.72	0.82	0.965
IGF + DMO + RFO (final)	0.97	0.97	0.67	0.79	0.9726

Architectural ablation as shown in Table. 8 demonstrated that multilayer perceptron and single-branch one-dimensional convolutional models underperform relative to contrastive Siamese training, while adding the Global Spatial and Channel Attention

block yields the highest AUC and best precision-recall behavior. Collectively, these results confirm that both hybrid feature selection and GSCA-enhanced Siamese representation learning are central to the model’s superior performance.

**Table 8. Ablation on network architecture components**

Model / Architecture	Accuracy	Precision	Recall	F1-score	ROC-AUC
Shallow MLP (1-2 layers)	0.93	0.94	0.63	0.75	0.92
Single-branch 1D-CNN	0.94	0.94	0.67	0.78	0.93
Siamese 1D-CNN (no GSCA)	0.96	0.96	0.70	0.81	0.96
Siamese 1D-CNN + GSCA (proposed)	0.97	0.97	0.67	0.79	0.9726

**5. State-of-Art Comparison**

Existing diabetes prediction models demonstrate varying performance across datasets and methodologies as shown in Table. 9. Prior works using smaller clinical datasets, such as those with 768 samples, report accuracies ranging from 84% to 95.76% [13-16], while studies using larger cohorts, including 2,942 samples, achieve up to 91.08% [17]. The most

recent model that was trained on 100,000 records achieved 96.09% accuracy [15], which underlines the advantage of large-scale data. Compared with these methods, the proposed hybrid IGF-DMO-RFO with Siamese GSCA architecture attains a superior 97% accuracy on the same dataset size, demonstrating improved discriminative power, better handling of class imbalance, and more robust feature representation.

**Table 9. State-of-Art Comparison**

Refer. No. /year	Size	Classes	Accuracy
[13]/2025	768	2 Classes Non-Diabetic, Diabetic	95.76%
[14]/2025	3546	2 Classes Non-Diabetic, Diabetes	73.01%
[15]/2025	100,000 records and 9 variables.	2 Classes: Diabetic, Non-Diabetic	96.09%
[16]/2025	768	2 Classes: Diabetic, Non-Diabetic	84%
[17]/2025	2,942	Class 1: Type 2 Diabetes Class 0: Non-diabetic	91.08%
Proposed Model	100,000	Class 0 - Non-Diabetic, Class 1 - Diabetic	97%

**6. DISCUSSION**

The results show that the effective combination of hybrid feature selection with Siamese embedding learning algorithm and GSCA attention algorithm provides an effective model for predicting diabetes. The IGF-DMOR pipeline eliminates noisy predictors and captures nonlinear interactions among features, as well as optimizes feature subsets, leading to compact and discriminative representations that lead to higher

training efficiency and less overfitting. Contrastive Siamese learning leads to further improvement of the class separation, while GSCA focuses on clinically relevant patterns of features and can be generalized better compared with traditional multilayer perceptrons or single-branch convolutional models. Overall, the proposed approach is a combination of an automated feature optimization strategy and robust representation learning, providing better performance

than ordinary machine-learning and deep-learning approaches.

## 7. CONCLUSION AND FUTURE WORK

The study introduces a reliable diabetes prediction model combining hybrid feature selection of IGF-DMO-RFO with Siamese one-dimensional CNN with GSCA attention. The 97% accuracy score with high discriminative performance and calibration means that the proposed model has achieved better performance than conventional machine learning and deep learning baselines. The hybrid feature selection approach plays a good role in suppressing noise and learning meaningful feature interaction, and the contrastive strategy, Siamese learning, generates well-separated embeddings under class imbalance. Furthermore, the GSCA module improves the representation quality in terms of clinically relevant factors such as HbA1c, glucose, BMI, age, and comorbidities, so that the final results of ROC-AUC, PR-AUC, and calibration performance are consistently high.

In conditions, the proposed model is adequate for clinical screening, risk stratification, and population-level health monitoring because of the efficient inference and compact architecture. However, the research is confined to one public dataset, and future research should involve multi-center validation to estimate the generalizability across various healthcare settings. Multi-modal data, such as longitudinal data from laboratory measurements, an imaging biomarker, or genomic information, may help further improve predictive performance.

## REFERENCES

1. Caballero-María, P., Caballero-Villarraso, J., Arenas-Montes, J., Díaz-Cáceres, A., Castañeda-Nieto, S., Alcalá-Díaz, J.F., Delgado-Lista, J., Rodríguez-Cantalejo, F., Pérez-Martínez, P., López-Miranda, J. and Camargo, A., 2025. Deep Learning Model Approach to Predict Diabetes Type 2 Based on Clinical, Biochemical, and Gut Microbiota Profiles. *Applied Sciences*, 15(4), p.2228.
2. Abousaber, I., Abdallah, H.F. and El-Ghaish, H., 2025. Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets. *Frontiers in Artificial Intelligence*, 7, p.1499530.
3. Lee, H., Hwang, S.H., Park, S., Choi, Y., Lee, S., Park, J., Son, Y., Kim, H.J., Kim, S., Oh, J. and Smith, L., 2025. Prediction model for type 2 diabetes mellitus and its association with mortality using machine learning in three independent cohorts from South Korea, Japan, and the UK: a model development and validation study. *EClinicalMedicine*, 80.
4. Lee, H., Park, M.B. and Won, Y.J., 2025. AI Machine Learning-Based Diabetes Prediction in Older Adults in South Korea: Cross-Sectional Analysis. *JMIR Formative Research*, 9(1), p.e57874.
5. Fan, Y., 2025. Diabetes diagnosis using a hybrid CNN LSTM MLP ensemble. *Scientific Reports*, 15(1), p.26765.
6. Tanabe, H., Sato, M., Miyake, A., Shimajiri, Y., Ojima, T., Narita, A., Saito, H., Tanaka, K., Masuzaki, H., Kazama, J.J. and Katagiri, H., 2024. Machine learning-based reproducible prediction of type 2 diabetes subtypes. *Diabetologia*, 67(11), pp.2446-2458.
7. Wee, B.F., Sivakumar, S., Lim, K.H., Wong, W.K. and Juwono, F.H., 2024. Diabetes detection based on machine learning and deep learning approaches. *Multimedia Tools and Applications*, 83(8), pp.24153-24185.
8. Kiran, M., Xie, Y., Anjum, N., Ball, G., Pierscionek, B. and Russell, D., 2025. Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature analysis. *Frontiers in Digital Health*, 7, p.1557467.
9. Kaliappan, J., Saravana Kumar, I.J., Sundaravelan, S., Anesh, T., Rithik, R.R., Singh, Y., Vera-Garcia, D.V., Himeur, Y., Mansoor, W., Atalla, S. and Srinivasan, K., 2024. Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets. *Frontiers in Artificial Intelligence*, 7, p.1421751.
10. El-Bashbishy, A.E.S. and El-Bakry, H.M., 2024. Pediatric diabetes prediction using deep learning. *Scientific Reports*, 14(1), p.4206.
11. Liu, H., Dong, S., Yang, H., Wang, L., Liu, J., Du, Y., Liu, J., Lyu, Z., Wang, Y., Jiang, L. and Yu, S., 2024. Comparing the accuracy of four machine learning models in predicting type 2 diabetes onset within the Chinese population: a retrospective study. *Journal of International Medical Research*, 52(6), p.03000605241253786.
12. Phongying, M. and Hiriole, S., 2023. Diabetes classification using machine learning techniques. *Computation*, 11(5), p.96.
13. Bhargale, K.B., Bhosale, S., Temkar, R., Adagale-Vairagar, S., Adagale, S.S., Mapari, R. and Tiwari, H., 2025. Diabetes Prediction from Clinical Data Using Deep Convolution Neural Network. *Mathematical Modelling of Engineering Problems*, 12(8).
14. Hageh, C.A., Henschel, A., Zhou, H., Zubelli, J., Nader, M., Chacar, S., Iakovidou, N., Hatzikirou, H., Abchee, A., O'Sullivan, S. and Zalloua, P.A., 2025. Improving T2D machine learning-based prediction accuracy with SNPs and younger age. *Computational and Structural Biotechnology Journal*.
15. Afolabi, S., Ajadi, N., Jimoh, A. and Adenekan, I., 2025. Predicting diabetes using supervised machine learning algorithms on E-health records. *Informatics and Health*, 2(1), pp.9-16.
16. Ghazizadeh, Y., Salehi, S. and Mirsaeid Ghazi, M., 2025. Machine learning-based diabetes prediction: A comprehensive study on predictive modeling and risk assessment. *J Clin Images Med Case Rep*, 6(5), p.3578.
17. Khan, S. and Shah, Z., 2025. Artificial intelligence-based diabetes risk prediction from longitudinal DXA bone measurements. *Scientific Reports*, 15(1), p.25706.
18. Lee, T.F., Chang, C.H., Chi, C.H., Liu, Y.H., Shao, J.C., Hsieh, Y.W., Yang, P.Y., Tseng, C.D., Chiu, C.L., Hu, Y.C. and Lin, Y.W., 2024. Utilizing radiomics and dosiomics with AI for precision prediction of radiation dermatitis in breast cancer patients. *BMC cancer*, 24(1), p.965.

19. Lugner, M., Rawshani, A., Helleryd, E. and Eliasson, B., 2024. Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data. *Scientific reports*, 14(1), p.2102.
20. Jiang, L., Xia, Z., Zhu, R., Gong, H., Wang, J., Li, J. and Wang, L., 2023. Diabetes risk prediction model based on community follow-up data using machine learning. *Preventive Medicine Reports*, 35, p.102358.
21. Feng, X., Cai, Y. and Xin, R., 2023. Optimizing diabetes classification with a machine learning-based framework. *BMC bioinformatics*, 24(1), p.428.
22. Aslan, M.F. and Sabanci, K., 2023. A novel proposal for deep learning-based diabetes prediction: converting clinical data to image data. *Diagnostics (Basel)* 13 (4): 796 [online]
23. Alanis, A.Y., Sanchez, O.D., Vaca-González, A. and Rangel-Heras, E., 2023. Intelligent classification and diagnosis of diabetes and impaired glucose tolerance using deep neural networks. *Mathematics*, 11(19), p.4065.
24. Alghamdi, T., 2023. Prediction of diabetes complications using computational intelligence techniques. *Applied Sciences*, 13(5), p.3030.
25. Machado-Fragua, M.D., Landré, B., Chen, M., Fayosse, A., Dugravot, A., Kivimaki, M., Sabia, S. and Singh-Manoux, A., 2022. Circulating serum metabolites as predictors of dementia: a machine learning approach in a 21-year follow-up of the Whitehall II cohort study. *BMC medicine*, 20(1), p.334.