http://jhmi.sums.ac.ir
Health Man & Info Sci
2026, 13(4), 295–307

HMIS

Health Management and Information Science

Original Article

Enhancing Clinical Reliability in Diabetes Mellitus Diagnosis: A Multi-Stage Deep Neural Network Framework

Parisa Eslami[1*], Zahra Karbasi[2]

[1]Health Human Resources Research Center, Shiraz University of Medical Sciences, Shiraz, Iran
[2] Department of Health Information Sciences, Faculty of Management and Medical Information Sciences, Kerman University of Medical Sciences, Kerman, Iran

Abstract:

Introduction: The framework comprises four phases: data acquisition, preprocessing, architectural modeling, and evaluation. To address data integrity, an unsupervised K-means clustering algorithm (k=3), evaluated via the Davies-Bouldin Index, was applied to standardized physiological attributes to identify and prune severe structural outliers (n=26). Subsequently, a feedforward Artificial Neural Network (ANN) and a Multi-Layer DL architecture incorporating ReLU activation and momentum optimization were evaluated. Validation was conducted using 10-fold Stratified Cross-Validation on both raw baseline data (N=768) and the refined cohort (N=742).

Methods: The framework comprises four phases: data acquisition, preprocessing, architectural modeling, and evaluation. To address data integrity, an unsupervised K-means clustering algorithm (k=3), evaluated via the Davies-Bouldin Index, was applied to standardized physiological attributes to identify and prune severe structural outliers (n=26). Subsequently, a feedforward Artificial Neural Network (ANN) and a Multi-Layer DL architecture incorporating ReLU activation and momentum optimization were evaluated. Validation was conducted using 10-fold Stratified Cross-Validation on both raw baseline data (N=768) and the refined cohort (N=742).

Results: Combining unsupervised outlier purification with multi-layer deep neural networks effectively mitigates clinical data noise and enhances model performance. This analytical pipeline provides a robust, scalable foundation for integration into Clinical Decision Support Systems (CDSS) to assist early diabetes screening and reduce the health and economic burdens of unmanaged diabetes risk.

Conclusion: Diabetes Mellitus; Deep Learning; PIMA Indian Dataset; Outlier Detection; Artificial Neural Networks

Keywords: Diabetes Mellitus; Deep Learning; PIMA Indian Dataset; Outlier Detection; Artificial Neural Networks

Article History:
Received: 01 August 2026
Accepted: 20 September 2026

Please cite this paper as:
Eslami P, Karbasi Z. Enhancing Clinical Reliability in Diabetes Mellitus Diagnosis: A Multi-Stage Deep Neural Network Framework. Health Man & Info Sci. 2026; 13(4): 295-307. doi: 10.30476/jhmi.2026.112718.1391

***Correspondence to:**
Parisa Eslami,
Health Human Resources Research Center, School of Health Management and Information Sciences, Shiraz University of Medical Sciences, Shiraz, Iran.
Email: ParisaEslamei@gmail.com

Introduction

Diabetes Mellitus (DM) is a critical non-communicable disease (NCD) and a primary global healthcare challenge. Characterized by chronic hyperglycemia resulting from defects in insulin secretion, insulin action, or both, DM leads to long-term damage and failure of various vital organs (1, 2).

Currently emphasized by the World Health Organization (WHO) as one of the four primary diseases, DM is central to many health policy

Copyright: ©Health Management and Information Science. This is an open-access article distributed under the terms of the Creative Commons Attribution-No Derivatives 4.0 International License. This license allows reusers to copy and distribute the material in any medium or format in unadapted form only, and only so long as attribution is given to the creator. The license allows for commercial use

initiatives. Beyond increasing the risk of cardiovascular diseases, peripheral neuropathy, renal failure, and visual impairments, the disease imposes a considerable economic burden on both patients and healthcare systems (3). Lifestyle changes associated with urbanization, rising obesity, decreased physical activity, poor diets, and increased life expectancy have contributed to a rapid and alarming growth in diabetes prevalence in recent decades (3, 4).

Reports published by the WHO indicate that over 500 million adults worldwide are afflicted with diabetes, a figure projected to increase dramatically to over 852 million in the coming decades. Furthermore, approximately half of those affected are unaware of their condition, with diagnosis often occurring only after the onset of chronic complications in advanced stages. This situation exponentially escalates treatment costs and limits the efficacy of clinical interventions (5-8).

Consequently, early screening and identification of at-risk individuals have emerged as among the most effective strategies for disease control and the prevention of long-term complications (9, 10). However, conventional early diagnostic tools often lack sufficient accuracy or require costly, time-consuming laboratory tests. Therefore, the development of accurate and reliable predictive models capable of identifying hidden patterns in the early stages of the disease has become fundamentally important (9, 11).

In Iran, diabetes is also considered one of the most common NCDs, with national studies indicating a prevalence of approximately 11% among adults (12). This accelerating trend, exacerbated by factors such as lifestyle changes, reduced physical activity, inadequate nutrition, and an aging population, represents a critical alarm for public health. The direct and indirect costs of diabetes in the country including expenses related to hospitalization, diabetic foot ulcers, dialysis, medication, and reduced workforce productivity, exert severe economic pressure on the healthcare system (13). Enhancing the accuracy of early diagnostic methods and prediction of onset is thus vital for national diabetes management (9).

The advancement of data science and the emergence of Machine Learning (ML) and Deep Learning (DL) have created vast new opportunities for clinical data analysis and disease prediction. These models can analyze large, complex datasets to identify nonlinear patterns and hidden relationships among clinical features, offering highly accurate predictions (14). In the context of diabetes prediction, leveraging simple clinical data, such as glucose levels, blood pressure, Body Mass Index (BMI), age, number of pregnancies, and other laboratory indices, allows for the construction of models that can estimate the probability of onset with acceptable accuracy, without requiring expensive or invasive procedures (15, 16).

The PIMA Indian Diabetes Dataset (PIDD) is a notable and widely used data source in this domain, having long served as a reference for evaluating the performance of diabetes diagnosis algorithms. Due to its comprehensiveness, high data quality, and inclusion of key clinical variables, it has facilitated the development of numerous models in international scholarly research (17).

Various studies have used classical ML techniques, including logistic regression (LR), Support Vector Machines (SVM), Random Forests (RF), Decision Trees (DT), Bayesian Networks (BN), and K-Nearest Neighbors (KNN), for diabetes prediction, yielding acceptable but not optimal performance. The limitations of these methods primarily stem from the inherent complexity of clinical data, the presence of non-linear relationships and multi-dimensional feature interactions, and issues such as imbalanced data and missing values (18-21).

In these circumstances, DL approaches with multi-layered architectures can automatically learn

features and extract complex patterns, offering significantly superior performance to traditional methods (22). Recent research has demonstrated that deep neural networks, hybrid models, and ensemble methods can boost prediction accuracy to 97%-99% and achieve high sensitivity in identifying at-risk individuals. Furthermore, techniques such as batch normalization, hyperparameter optimization, and data augmentation can improve model stability and generalizability (22-24).

Despite these advancements, past studies have yet to fully address various aspects of optimizing deep architectures, selecting optimal features, managing imbalanced data, and designing highly generalizable models. Differences in demographics, the diversity of clinical features, limitations in available data, and the necessity for rigorous evaluation under diverse conditions demand more comprehensive, precise research based on novel designs. Therefore, developing a DL-based model that provides an effective tool for diabetes prediction through appropriate pre-processing methods, network architecture tuning, and rigorous performance evaluation is a scientific and practical imperative.

The present study aims to develop an accurate, stable, and practical model for early screening of at-risk individuals by leveraging Artificial Neural Networks (ANNs) and DL algorithms. This will be accomplished through model-structure optimization and rigorous evaluation using the PIDD to develop an accurate, stable, and practical model for initial screening of at-risk individuals. The outcome of this research is expected to be a significant step toward reducing national healthcare costs, improving the quality of care, mitigating the incidence of chronic complications, and supporting data-driven clinical decision-making. From an applied perspective, this research can also play an essential role in improving health policies and enhancing the quality of life for patients and at-risk individuals.

Materials and Methods

The protocol of this study was approved by the Ethics Committee of Shiraz University of Medical Sciences (SUMS) under the approval code (IR.SUMS.NUMIMG.REC.1405.051).

The research pipeline is partitioned into four sequential phases: Data Acquisition, Preprocessing, Model Architecture Design, and Performance Evaluation. The specific procedures implemented within each stage are detailed in the following sections.

Data Description

The statistical population used in this research is the PIDD, a public, well-established, widely-cited dataset originally sourced from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (17).

The dataset comprises 768 cases characterized by eight distinct features, including patient demographics, key laboratory measurements, and clinical indicators. Of the total samples, 268 cases are classified as positive (at risk/diabetic), while 500 cases are classified as negative (healthy). The data originates from a population residing in the Arizona region of the United States (17). The dataset variables are described in detail in Table 1.

Data Preprocessing

Data Integrity Assessment

The initial phase of data preprocessing focused on ensuring the dataset's completeness and integrity, a critical prerequisite for the validity of all subsequent modeling and analysis. This step involved detecting and addressing common data quality issues, specifically missing values and duplicate records.

Outlier Detection via K-means Clustering

In this research, clustering-based techniques are employed for outlier detection. Specifically, the K-means clustering algorithm is utilized. K-means is an iterative partitioning method that begins by randomly selecting k initial points as

cluster centroids. Data points are subsequently assigned to the nearest cluster based on their proximity. The centroids are then recalculated by averaging the data points within each cluster, and the assignment process repeats until convergence is reached—that is, when no further changes occur in the cluster assignments.

The primary objective of an effective clustering algorithm is to minimize intra-cluster distance (maximizing cohesion) and maximize inter-cluster distance (maximizing separation). Consequently, cluster density and cluster separation are the two critical factors in evaluating performance.

To evaluate the quality of the clustering and determine the optimal configuration, the Davies-Bouldin (DB) Index is applied (25). This index is defined based on two fundamental factors:

Separation (S): Defined as the Euclidean distance between the centroids of two clusters (Equation 1).

$$S_{i,j} = \text{dist}(c_i, c_j) = \sqrt{\sum_{k=1}^{n}(c_{i,k} - c_{j,k})} \quad (1$$

Cohesion (T): Calculated as the average distance between each point x within a cluster and its corresponding centroid c (Equation 2).

$$T_i = \frac{1}{N_i}\sum_{x \in C_i} \text{dist}(x, c_i) \quad (2)$$

The general formula for the Davies-Bouldin Index is expressed as (Equation 3):

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{T_i + T_j}{S_{i,j}}\right) \quad (3)$$

The clustering process and evaluation were implemented using the Cluster Distance Performance operator within the RapidMiner platform. The DB criterion was selected for evaluation. This index provides a normalized value, with a score closer to zero indicating superior clustering performance. To ensure the transparency and reproducibility of the outlier detection process, a standardized configuration was established. These parameters were selected to maximize the sensitivity of the K-means algorithm to structural anomalies within the PIDD feature space. The implementation parameters are described in Table 2.

Table 1. Description of the selected features used in PIDD

Feature	Description	Measurement Unit	Type
Pregnancies	Number of times pregnant	Count	Discrete
Plasma glucose concentration	Plasma glucose concentration (2 hours in an oral glucose tolerance test)	mg/dL	Continuous
Blood Pressure	Diastolic blood pressure	mm Hg	Continuous
Skin Thickness	Triceps skin fold thickness	mm	Continuous
Insulin	2-hour serum insulin	mu U/ml	Continuous
BMI	Body Mass Index	kg/m²	Continuous
Pedigree	Diabetes Pedigree Function	Score	Continuous
Age	Years of age	Years	Discrete
Outcome Variable	Diabetic status (0 = No, 1 = Yes)	Binary	Categorical

Table 2. K-means Clustering Configuration and Parameters

Parameter	Configuration	Description
Cluster Count (k)	3 (optimized)	Determined via empirical testing of k in. Selection based on minimizing the Davies-Bouldin Index
Target Variable	Excluded	The label variable was removed to ensure a purely unsupervised learning approach.
Data Type	Numeric	No transformation required as all attributes were natively numerical.

Distance Measure	Squared Euclidean	Utilized via Bregman Divergences for precise proximity calculation.
Optimization Goal	Centroid-based	The evaluator was used to minimize intra-cluster distance and identify outliers.

Classification Algorithms

In this study, two distinct neural architectures were implemented to evaluate predictive performance on the preprocessed data: an ANN and a DL model.

The ANN was selected to assess the model's ability to handle dependencies among clinical features using a single-layer architecture. A key feature of this implementation is the inclusion of a Learning Rate Decay coefficient. This mechanism allows for larger weight adjustments at the start of training to explore the error surface, followed by finer refinements in later stages to prevent overfitting and ensure stable convergence.

The DL model was designed as a multi-layered architecture to capture higher-level abstractions. This model uses the Rectified Linear Unit (ReLU) activation function, which is chosen to mitigate the vanishing gradient problem common in deeper networks. Furthermore, feature importance was integrated into the input layer to prioritize variables with the highest clinical significance, and a high momentum coefficient was applied to preserve the gradient direction in deeper layers.

The specific hyperparameters and architectural configurations for both models are summarized in Table 3.

Table 3. Comparative Hyperparameter Configuration for ANN and D L Models

Parameter	ANN Model	DL Model
Architecture	1 Hidden Layer (8 Inputs)	2 Hidden Layers (20 Inputs)
Activation Function	Sigmoid / Tanh	Rectifier (ReLU)
Training Cycles / Epochs	500 Cycles	10 Epochs
Learning Rate	0.01	Adaptive
Momentum	0.7	0.99
Learning Rate Decay	Enabled	N/A
Input Feature Selection	Standard	Importance-based

Model Evaluation

To rigorously assess the performance of the proposed classification models, several standard metrics derived from the Confusion Matrix were utilized (26). As shown in Table 4, this matrix provides a detailed breakdown of the classifier's performance by comparing predicted labels against the dataset's ground truth. For a binary classification task involving classes C_1 (positive) and C_2 (negative), the matrix is composed of four fundamental outcomes:

Table 4. The Confusion Matrix

Actual Prediction	Predicted Healthy	Predicted Diabetic
Healthy	TN	FP
Diabetic	FN	TP

True Positive (TP): Instances where the actual class is positive, and the model correctly predicts them as positive.

False Positive (FP): Instances where the actual class is negative, but the model incorrectly predicts them as positive (Type I error).

True Negative (TN): Instances where the actual class is negative, and the model correctly predicts them as negative.

False Negative (FN): Instances where the actual class is positive, but the model incorrectly predicts them as negative (Type II error)

Accuracy

Accuracy represents the proportion of total instances that were correctly identified by the classifier. It is calculated using the following equation (Equation 4)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4$$

Sensitivity

Sensitivity measures the classifier's ability to correctly identify patients with the disease. In a clinical context, it represents the ratio of correct positive diagnoses to the total number of individuals who truly have the condition:(Equation 5)

$$Sensitivity = \frac{TP}{TP + FN} \quad (5)$$

Specificity

Specificity measures the model's ability to correctly identify healthy individuals. It represents the proportion of actual negative cases that were correctly identified as such: (Equation 6)

$$Specificity = \frac{TN}{TN + FP} \quad (6)$$

Area Under the ROC Curve (AUC-ROC)

The ROC curve is a two-dimensional visualization of a classifier's performance. The vertical axis represents the TP Rate (Sensitivity), while the horizontal axis represents the FP Rate (1 - Specificity). The AUC is a critical metric for assessing the overall efficiency of the algorithm; a higher AUC, closer to 1.0, indicates greater discriminative power.

Model Validation

In supervised learning problems, specifically classification tasks, the dataset is typically partitioned into training and testing subsets. While a simple validation approach can be employed, this study utilizes k-fold Cross-Validation to ensure a more robust and comprehensive evaluation. This method ensures that every data point in the entire dataset is used for both training and testing at least once, thereby providing a more reliable estimation of model performance.

In this procedure, the original dataset is partitioned into k equal-sized subsamples, namely folds. Of the k folds, a single fold is retained as the validation data for testing the model, while the remaining k-1 folds are used as training data. This process is repeated k times, with each of the k folds used exactly once as the validation data. The results from the k iterations are then averaged to produce a single, final performance estimate.

For the implementation in RapidMiner, this cross-validation framework was applied. Furthermore, a Stratified Sampling technique was selected to ensure that each fold maintains the same class proportions as the original dataset, thereby mitigating class imbalance and improving the stability of the classification results.

Results

Results of Data Preprocessing

Upon execution of the clustering algorithm, a new categorical attribute labeled "Cluster" was appended to the dataset, identifying the membership of each record. The distribution of data points across the resulting clusters is illustrated in Table 5.

Table 5. Distribution of Data Instances Across the Identified Clusters

Cluster Model	Number of Items
Cluster 0	511
Cluster 1	26
Cluster 2	231
Total number of items	768

The DB Index was employed as the primary internal validity metric to determine the optimal number of clusters. The comparative results for k=2 and k=3 are 0.089 and 0.079, respectively.

As established in the literature, a lower DB Index indicates superior cluster separation and internal cohesion. Based on the experimental results, the configuration with k=3 yielded a significantly

lower index, suggesting higher optimization. Consequently, the subsequent analysis is conducted using three clusters.

To interpret the clinical characteristics of each group, a Centroid Table analysis was performed as illustrated in Table 5.

Table 6. Centroid Table analysis

Attribute	Cluster 0	Cluster 1	Cluster 2
Pregnancies	3.912	3.538	3.732
Plasma glucose concentration	113.703	161.846	132.195
Diastolic blood pressure	67.685	70.615	72.078
Triceps skin fold thickness	15.405	33.385	30.442
Serum Insulin	16.832	497	172.134
Body Mass Index	30.765	35.550	34.308
Diabetes Pedigree Function	0.434	0.601	0.540
Age	33.481	33.692	32.658

The profile of each cluster is defined as follows:

- Cluster 2 (Diabetic Group): This cluster is characterized by high mean Insulin levels (172.13), alongside elevated skin thickness, blood pressure, and BMI. Furthermore, this group shows a higher Diabetes Pedigree Function (hereditary factor) than others. It contains 231 instances primarily representing Type 2 Diabetic patients.
- Cluster 0 (Healthy Group): This cluster represents healthy individuals, characterized by lower mean Insulin, BMI, Glucose, and blood pressure compared to the other two groups. This is the largest segment, containing 511 instances.
- Cluster 1 (Outlier Group): This cluster contains only 26 instances and exhibits abnormally high values for Insulin, blood pressure, and weight. These values represent mathematical and clinical anomalies that deviate significantly from the general population and were targeted for removal.

Scatter plots were generated to visualize the spatial distribution of the clusters, as illustrated in Figure 1. The vertical axis represents 2-Hour Serum Insulin, while the horizontal axis represents Plasma Glucose Concentration. Following the identification of Cluster 1 as an outlier group, these 26 records were excluded to refine the training set.

The scatter plot clearly demonstrates that Cluster 1 resides at a significant distance from the core data distribution. While some overlap is observed in Glucose levels, the separation is most distinct along the Insulin axis.

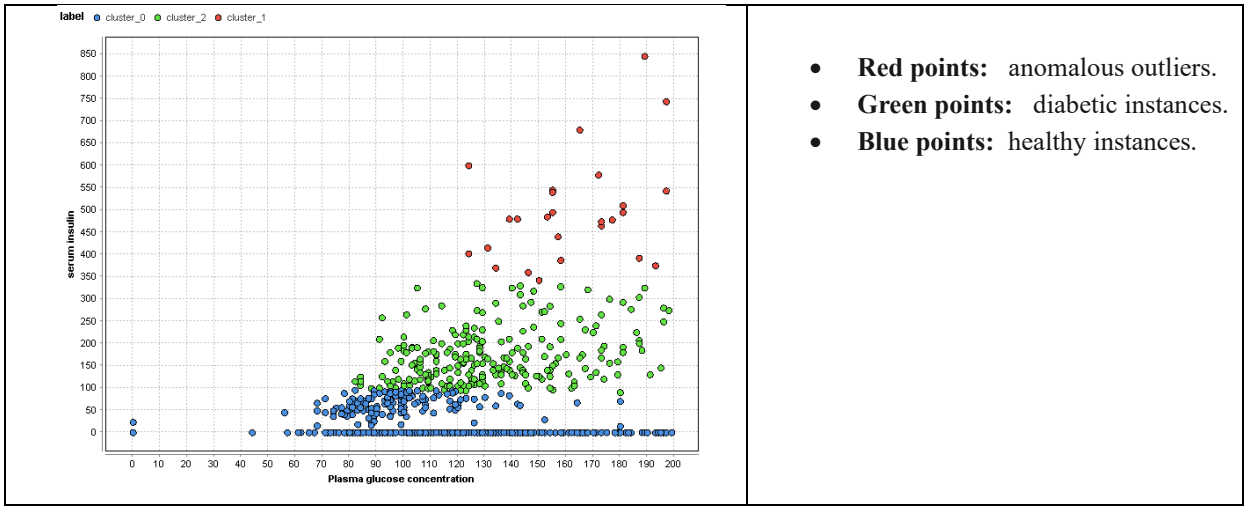
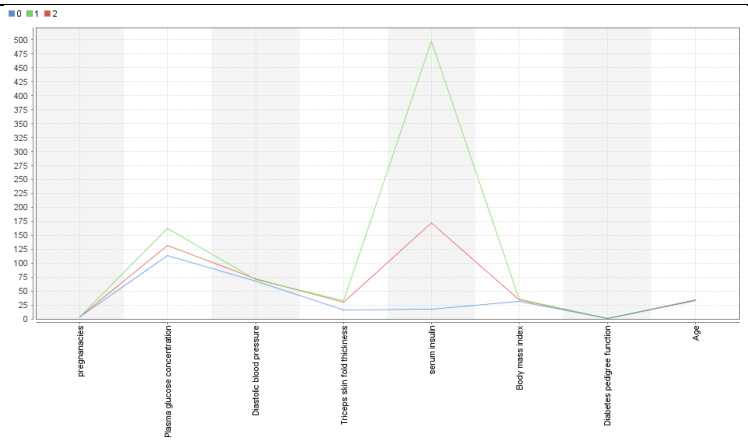


Figure 1. Spatial distribution of clinical clusters following K-means analysis, highlighting identified outliers against the refined diagnostic classes

The most significant insights are provided by the Parallel Threshold (Deviation) Plot as illustrated in Figure 2. Each colored line represents the mean score of a cluster across all features.

The deviation analysis confirms that while all features contribute to the model, Serum Insulin and Plasma Glucose are the most critical discriminative variables for cluster separation, with Insulin exhibiting the highest variance across groups.



- **Blue line:** the healthy cohort
- **Red line:** the diabetic cohort.
- **Green line:** the outlier cohort

Figure 2. Parallel coordinate plot illustrating the mean feature profiles and cluster deviations across all physiological variables.

Results of Model Evaluation and Validation

In this section, we evaluate the performance of the ANN and DL models. The analysis is conducted in two distinct stages: first, utilizing the raw dataset as a baseline, and second, applying clustering-based outlier removal to assess the impact of data preprocessing on model efficacy. As illustrated in Figure 3, the outlier detection process resulted in the exclusion of 26 anomalous samples. The remaining observations were partitioned into two clusters to define the new target labels: Cluster 2 (Diabetic class, n=231) and Cluster 1 (Healthy class, n=511). This clustering output was subsequently designated as the ground-truth variable (target label) for the classification phase.

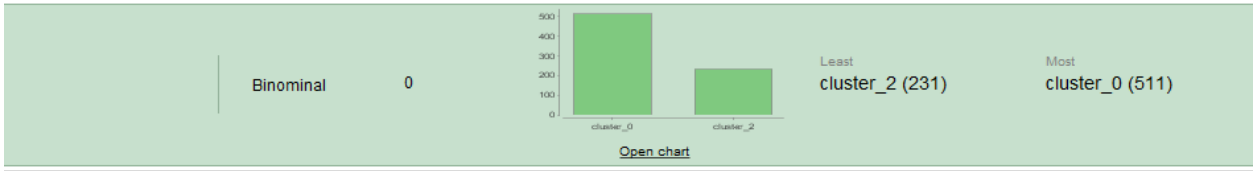


Figure 3. Classification of the refined target variable based on clustering results.

The classification performance for both models, with and without preprocessing, is summarized in Table7.

Table 7. Comparative Performance Metrics of ANN and Deep Learning Models

Model	Configuration	Accuracy	Sensitivity	Specificity	AUC
ANN	Baseline (Raw Data)	65.11%	100.00%	0.00%	0.802
DL	Baseline (Raw Data)	76.17%	90.00%	50.37%	0.833
ANN	Post-Clustering	71.70%	9.02%	100.00%	0.993
DL	Post-Clustering	98.65%	97.83%	99.02%	0.990

A primary observation from Table 2 is that preprocessing significantly enhances model accuracy. By clustering the data into two distinct groups—Cluster 2 (Diabetic, n=231) and Cluster 1 (Non-diabetic, n=511)—and removing 26 anomalous outliers, the models achieved a more refined classification. Notably, the DL model exhibited an average performance increase of approximately 27% across all evaluation metrics.

The results indicate that DL consistently outperforms the standard ANN architecture in both experimental stages. This improvement is attributed to the deep architecture and its ability to process a high number of inputs, which allowed it to capture complex patterns within the PIDD more effectively.

Furthermore, while the ANN achieved 100% sensitivity in certain cases, it suffered from severe class imbalance issues, as evidenced by the 0% specificity in the raw stage. In contrast, the DL model demonstrated robust stability, maintaining exceptionally high accuracy for both classes simultaneously, suggesting that it is less susceptible to dataset imbalances.

To further investigate the classification behavior, the confusion matrices for both models post-clustering were analyzed.

The ANN post-clustering displayed a significant bias toward the negative (healthy) class. As shown in Table 8, the model misclassified 210 diabetic cases as healthy (FN), indicating that the classifier failed to generalize the positive class effectively despite a high overall AUC.

Table 8. Confusion Matrix for the Proposed ANN Model

Actual Prediction	Predicted: Healthy	Predicted: Diabetic
Healthy	511	0
Diabetic	210	21

The DL model achieved near-perfect classification, exhibiting high precision across both categories. As shown in Table 9, with only 10 total errors across 742 samples (5 FP and 5 FN), the DL model demonstrates a balanced and reliable predictive capability.

Table 9. Confusion Matrix for the Proposed Deep Learning Model.

Actual Prediction	Predicted: Healthy	Predicted: Diabetic
Healthy	505	5
Diabetic	5	226

The superior discriminative power of the DL model is also reflected in the ROC curves (Figures 4 to 7), where the DL model maintains a near-ideal curve post-clustering.

Discussion

This research proposed a hybrid analytical framework for diabetes prediction by integrating unsupervised K-means clustering with a deep neural architecture. The core of this work involved a two-stage process: first, the identification and removal of 26 clinical anomalies to refine the stochastic variance within the dataset; and second, the implementation of a multi-layered DL model. By transitioning from the statistically inconsistent original labels to physiologically homogeneous clusters,

the proposed model achieved a peak accuracy of 98.65%, a sensitivity of 97.83%, and a specificity of 99.02%.

A central finding of this research is that model performance on raw clinical data is inherently limited by spurious data points and biological anomalies. The initial baseline testing yielded an accuracy of only 76.17% for the DL model and a biased 65.11% for the ANN. This significant performance gap highlights the necessity of extensive pre-processing, a sentiment echoed by Salih et al. (2024) and Kaur & Kumari (2019), who utilized PCA and Boruta algorithms respectively to refine feature subsets (19, 27).

However, rather than focusing solely on dimensionality reduction, our approach utilized K-means clustering to isolate 26 extreme physiological outliers. Following this exclusion and the subsequent re-labeling of the remaining 742 samples into physiologically consistent groups, we observed a 27% average performance increase across all metrics. This confirms the premise established by Pandya (2023), who noted that addressing dataset-specific irregularities specifically through balancing techniques like ADASYN, is vital for reliability (28). Our study extends this logic by proving that clustering-based cleaning offers a more stable foundation for deep architectures than raw data manipulation alone.

When benchmarked against recent studies utilizing the PIDD, the proposed DL model exhibits superior predictive power. While ensemble methods, such as the stacking approach by Reza et al. (2024) and the IoMT-focused RF by Chang et al. (2022), provided robust frameworks, their accuracies remained below the 80% threshold (29, 30). Even modern sequential models like the RFE-GRU proposed by Shams et al. (2025), which effectively captures feature interdependencies, achieved a lower accuracy of 90.70% (31).

The most comparable performance in current literature is found in the hybrid framework of Chellappan & Rajaguru (2025), which achieved 98.57% accuracy using a complex combination of ABC-PSO feature extraction and metaheuristic selection (32). Our proposed method achieved a slightly higher accuracy of 98.65% through a more streamlined pipeline. This parity suggests that rigorous data purification can be as effective as high-complexity metaheuristic selection, potentially offering lower computational overhead while capturing the complex, non-linear relationships within PIDD variables.

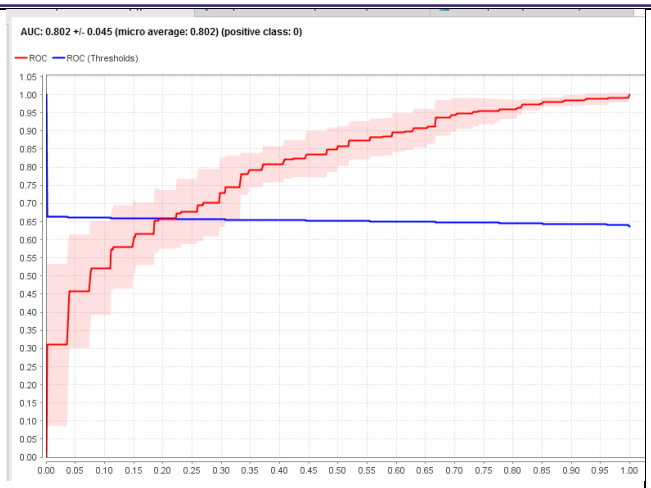


Figure 4-ROC Curve for the ANN using the raw dataset

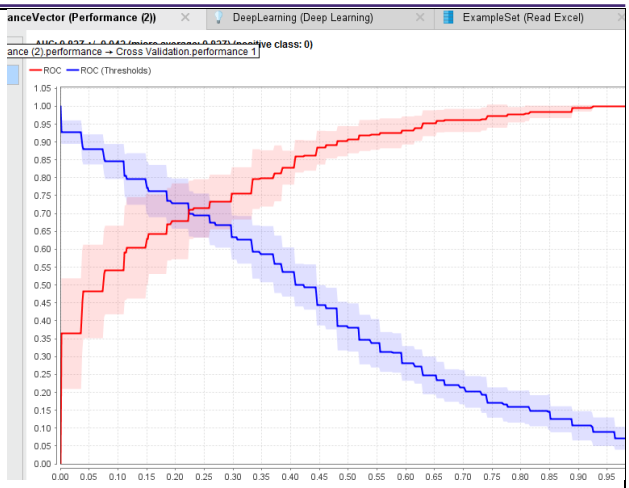


Figure 5-ROC Curve for the DL model using the raw dataset.

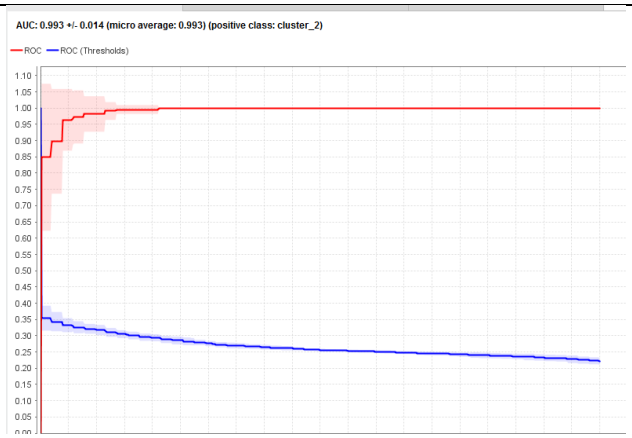


Figure 6-ROC Curve for the ANN post-outlier removal

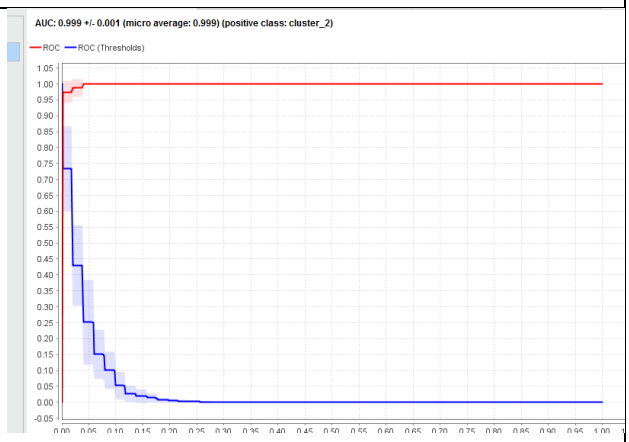


Figure 7-ROC Curve for the DL post-outlier removal

A significant challenge in diabetes prediction is maintaining a balance between Sensitivity and Specificity. Our analysis of the ANN confusion matrix revealed a severe systemic **bias** toward the majority class, misclassifying 210 diabetic instances as healthy individual. Such a failure rate is unacceptable in clinical diagnostics. In contrast, the DL model proved highly resilient to class distribution imbalances, achieving nearly symmetrical accuracy across both classes (97.83% for diabetics and 99.02% for healthy individuals).

This stability mirrors the objectives of Chang et al. (2022), who prioritized building clinical trust through interpretable models (30). Furthermore, our analysis confirms that Glucose, BMI, and Insulin are the most critical discriminative features, a finding consistent with Shams et al. (2025) **(**31**)**. By focusing on the spatial separation provided by K-means to identify "latent" patient profiles, we ensure that the high performance is grounded in clinical reality. The DL model's ability to limit errors to only 10 out of 742 cases underscores its potential as a reliable tool for automated e-diagnosis systems.

Despite the high predictive accuracy achieved by the proposed hybrid framework, several limitations must be acknowledged. First, the study relies exclusively on the PIDD. Therefore, the generalizability of our model to diverse global populations, who may exhibit different

physiological norms and diabetes risk profiles, requires further validation using multi-center and real-world clinical data. Second, while our model demonstrates excellent predictive performance, it operates as a black box in certain aspects; integrating Explainable AI (XAI) techniques, such as SHAP or LIME, would be essential to translate these computational predictions into actionable clinical insights for healthcare professionals. Furthermore, the model is currently designed for static data analysis and does not account for longitudinal progression of the disease, which remains a crucial direction for real-time monitoring systems.

Conclusion

This research demonstrates that the integration of unsupervised clustering with a deep supervised architecture constitutes a robust and highly effective methodology for automated diabetes screening. By bridging the gap between rigorous data purification and advanced neural modeling, this study establishes a new performance benchmark for the PIDD. Achieving an accuracy of 98.65% and an AUC of 0.99, our proposed framework effectively mitigates the stochastic variance inherent in clinical records, confirming that high-quality data processing is as critical as model architecture. These findings suggest that such an analytical pipeline offers a scalable solution for real-world healthcare infrastructures, facilitating data-driven clinical decision-making. Ultimately, by enabling the precise and early identification of high-risk patient profiles, this approach holds the potential to significantly alleviate the substantial economic and health burdens associated with undiagnosed diabetes.

Declarations

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics Approval

As this study involves the analysis of clinical data retrieved from PIMA Indian dataset, all procedures were conducted in accordance with the Declaration of Helsinki. No further administrative permissions were required for the secondary analysis of this publicly available data.

Conflict of Interest

The authors declare that they have no competing interests that could have influenced the work reported in this paper.

References:

1. Blair M. Diabetes mellitus review. Urol Nurs. 2016;36(1):27.
2. Bastaki S. Diabetes mellitus and its treatment. Int J Diabetes Metab. 2005;13(3):111-34.
3. Congdon P. Obesity and urban environments. Int J Environ Res Public Health. 2019;16(3):464.
4. Den Braver N, Lakerveld J, Rutters F, Schoonmade L, Brug J, Beulens J. Built environmental characteristics and diabetes: a systematic review and meta-analysis. BMC Med. 2018;16(1):12.
5. Wang Y, Lin T, Lu J, He W, Chen H, Wen T, et al. Trends and analysis of risk factor differences in the global burden of chronic kidney disease due to type 2 diabetes from 1990 to 2021: a population-based study. Diabetes Obes Metab. 2025;27(4):1902-19.
6. Haider R. The global burden of diabetes. Int J Clin Med Case Rep. 2024.
7. Abid HMU, Hanif M, Afzal MNB, Azeem M, e Shahwar D, Mahmood K. Global prevalence and mortality of type-2 diabetes from 1990 to 2019, with future projections to 2023 and 2050: a systematic review. Glob Drug Des Dev Rev. 2024;9(1):1-10.
8. Atique R, Rana T, Fatima HR, Shah RR, Sharif J, Naveed A, et al. Diabetes mellitus disclosure: a comprehensive review on a global issue. Glob J Multidiscip Sci Art 84-89 (1).
9. Hossain MJ, Al-Mamun M, Islam MR. Diabetes mellitus, the fastest growing global public health concern: early detection should be focused. Health Sci Rep. 2024;7(3):e2004.
10. Facciolà A, Visalli G, D'Andrea G, Varvarà M, Santoro G, Cuffari R, et al. Prevention of cardiovascular diseases and diabetes:

importance of a screening program for the early detection of risk conditions in a target population. J Prev Med Hyg. 2022;62(4):E934.
11. Nazirun NNN, Wahab AA, Selamat A, Fujita H, Krejcar O, Kuca K, et al. Prediction models for type 2 diabetes progression: a systematic review. IEEE Access. 2024;12:161595-619.
12. Shahraz S, Moghaddam SS, Azmin M, Peykari N, Yoosefi M, Mohebi F, et al. Prevalence of diabetes and prediabetes, and achievements in diabetes control in Iran; the results of the STEPS of 2016. Arch Iran Med. 2022;25(9):591.
13. Aminorroaya A, Fattahi N, Azadnajafabad S, Mohammadi E, Jamshidi K, Rouhifard Khalilabad M, et al. Burden of non-communicable diseases in Iran: past, present, and future. J Diabetes Metab Disord. 2024;23(2):1793-9.
14. Rahman A, Debnath T, Kundu D, Khan MSI, Aishi AA, Sazzad S, et al. Machine learning and deep learning-based approach in smart healthcare: recent advances, applications, challenges and opportunities. AIMS Public Health. 2024;11(1):58.
15. Zhou H, Rahman S, Angelova M, Bruce CR, Karmakar C. A robust and generalized framework in diabetes classification across heterogeneous environments. Comput Biol Med. 2025;186:109720.
16. Abousaber I, Abdallah HF, El-Ghaish H. Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets. Front Artif Intell. 2025;7:1499530.
17. Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS, editors. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proc Annu Symp Comput Appl Med Care. 1988.
18. Makroum MA, Adda M, Bouzouane A, Ibrahim H. Machine learning and smart devices for diabetes management: systematic review. Sensors (Basel). 2022;22(5):1843.
19. Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. Appl Comput Inform. 2022;18(1-2):90-100.
20. Fregoso-Aparicio L, Noguez J, Montesinos L, García-García JA. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. Diabetol Metab Syndr. 2021;13(1):148.
21. De Silva K, Lee WK, Forbes A, Demmer RT, Barton C, Enticott J. Use and performance of machine learning models for type 2 diabetes prediction in community settings: a systematic review and meta-analysis. Int J Med Inform. 2020;143:104268.
22. Zhu T, Li K, Herrero P, Georgiou P. Deep learning for diabetes: a systematic review. IEEE J Biomed Health Inform. 2020;25(7):2744-57.
23. Kiran M, Xie Y, Anjum N, Ball G, Pierscionek B, Russell D. Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature analysis. Front Digit Health. 2025;7:1557467.
24. Wee BF, Sivakumar S, Lim KH, Wong WK, Juwono FH. Diabetes detection based on machine learning and deep learning approaches. Multimed Tools Appl. 2024;83(8):24153-85.
25. Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell. 1979;(2):224-7.
26. Rashidi HH, Albahra S, Robertson S, Tran NK, Hu B. Common statistical concepts in the supervised machine learning arena. Front Oncol. 2023;13:1130229.
27. Salih MS, Ibrahim RK, Zeebaree S, Asaad D, Zebari LM, Abdulkareem NM. Diabetic prediction based on machine learning using PIMA Indian dataset. Commun Appl Nonlinear Anal. 2024;31(5s):138-56.
28. Pandya S. Predicting diabetes mellitus in healthcare: a comparative analysis of machine learning algorithms. Int J Curr Eng Technol. 2023;13(6).
29. Reza MS, Amin R, Yasmin R, Kulsum W, Ruhi S. Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data. Heliyon. 2024;10(2).
30. Chang V, Bailey J, Xu QA, Sun Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Comput Appl. 2023;35(22):16157-73.
31. Shams MY, Tarek Z, Elshewey AM. A novel RFE-GRU model for diabetes classification using PIMA Indian dataset. Sci Rep. 2025;15(1):982.
32. Chellappan D, Rajaguru H. Generalizability of machine learning models for diabetes detection: a study with Nordic islet transplant and PIMA datasets. Sci Rep. 2025;15(1):4479.