Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Leakage-Free and Statistically Validated Machine Learning Evaluation Framework for Diabetes Risk Prediction Across Multiple Datasets

Amresha Soomro
Altinbas University
Oguz Ata
oguz.ata@atlas.edu.tr
Atlas University

Research Article

Keywords: Diabetes Prediction, Machine Learning, Data leakage, Statistical model comparison

Posted Date: June 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9678084/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Leakage-Free and Statistically Validated Machine Learning Evaluation Framework for Diabetes Risk Prediction Across Multiple Datasets

Amresha Soomro[1] and Oguz Ata[2*†]

[1]Science and Engineering, Altinbas University, Istanbul, Turkey.
[2*]Science and Engineering, Atlas University, Istanbul, Turkey.

*Corresponding author(s). E-mail(s): oguz.ata@atlas.edu.tr;
Contributing authors: amreshasoomro@gmail.com;
†These authors contributed equally to this work.

Abstract

Machine learning models for diabetes classification frequently encounter issues with data leakage and inconsistent validation methodologies, rendering performance evaluations unreliable. This study presents a machine learning framework that mitigates data leakage and integrates fold-contained pre-processing with multi-seed stratified 10-fold cross-validation to guarantee fair evaluation. We examined six models: Logistic Regression, Random Forest, Feedforward Neural Network, Backpropagation Neural Network, General Regression Neural Network (GRNN), and Artificial Bee Colony Optimised Feedforward Neural Network.

We conducted analyses on the Pima Indians Diabetes Dataset and the Early-Stage Diabetes Risk Dataset. In Dataset 1, models attained ROC–AUC values ranging from 0.76 to 0.84, with the GRNN model demonstrating the highest performance (AUC = 0.837 ± 0.002). Logistic Regression (AUC = 0.832) and Random Forest (AUC = 0.829) produced similar outcomes. Dataset 2 exhibited superior performance, with GRNN achieving an AUC of 0.998, followed by ABC-FFNN (AUC = 0.991) and FFNN (AUC = 0.982). The calibration results indicated that GRNN excelled in probability estimation.

Statistical analysis confirmed notable performance differences among selected model pairs. The proposed framework enhances the reliability and reproducibility of diabetes classification research, facilitating a more comprehensive machine learning evaluation in healthcare environments.

Keywords: Diabetes Prediction, Machine Learning, Data leakage, Statistical model comparison.

1 Introduction

Diabetes mellitus is a major public health concern worldwide because it is becoming more common and can lead to serious long-term problems with the heart, kidneys, and metabolism. It's important to find risky people early on so that their health can be kept in check and problems can be avoided [1]. Recently, researchers have been looking into how machine learning can be used to predict diabetes based on clinical data that is collected on a regular basis. Neural network models have shown their usefulness in sorting data on benchmark datasets like the Pima Indian Diabetes Dataset [2],[3]. People are worried about how reliable and reproducible medical machine learning results are, especially when experimental pipelines allow data leakage, selective reporting, or not enough of a difference between model tuning and testing [4]. These kinds of methodological problems can make performance estimates too high and make predictive models less useful in the real world. This study presents a machine learning evaluation framework that is impervious to leaks and evaluates the effectiveness of foundational machine learning models and neural network architectures in systematically forecasting diabetes risk. The study investigates models that utilise two frequently used public datasets: the PIMA Indian Diabetes dataset and the Early-Stage Diabetes Risk dataset.

The next part looks at research that uses baseline machine learning models and neural networks to predict diabetes. This summary talks about the latest changes in predictive modelling and gives a list of some of the most common algorithms and how to test them.

It also talks about the problems with the methods used in earlier studies. This review looks at these studies to come up with the theoretical and methodological framework for the current research:

Machine learning has changed the way diabetes is managed, going from static classification to dynamic, cloud-based decision support and remote monitoring systems [3],[4],[5]. These frameworks focus on scalability and real-time risk assessment, and they often use ML fusion techniques to combine predictive power from different clinical settings [4],[5]. However, even though the technical infrastructure for deployment has improved, the methods used often put raw predictive performance ahead of the strict validation needed for clinical safety. This focus has made the field more scattered, with reports of high accuracy but no proof that the models work well in the real world.

A lot of recent research has been about making input data more reliable by using advanced pre-processing and feature engineering. For example, K-Nearest Neighbour imputation and tri-ensemble models have been used to deal with the common problem of missing medical records [6]. Other researchers have used gender-specific cohorts or community follow-up data to find unique risk patterns [7],[8]. A growing number of studies have combined "black-box" models with clinical trust to close the gap. Explainable AI and SHAP-based interpretability can help find clinically important features [9],[10],[11],[12]. Even though there have been advancements in how data is handled and how clear it is, there is still a major methodological gap: data leakage and model calibration are not always taken into account.

The "reliability gap" is most apparent in the reporting of performance metrics across standardised datasets such as PIMA [8],[13],[14]. Many studies are exceptionally

accurate, but a systematic review of 40 progression models found that less than 2% of published papers report important clinical metrics like calibration scores (MCE/Brier Score). This implies that most of the models which are currently available are best at telling the difference between things (accuracy) but don't make sure that their predicted probabilities match actual clinical risk [15].

Early diabetes prediction is essential for avoiding serious problems like kidney failure and heart disease, which is why it is a major focus of ML applications [10],[16]. This literature review discovers that leakage-free validation, calibration analysis, and formal statistical testing are not always used, even though ML is widely used because it can map non-linear physiological patterns. "Optimal bias" and a lack of statistical rigour slow down the move from research to clinical practice without these. The current study specifically addresses this fragmentation by providing a machine learning framework that doesn't leak and systematically tests six baseline and neural network models on two diabetes datasets using multi-seed stratified 10-fold cross-validation with fold-contained preprocessing. This work sets a new standard for openness and reproducibility in the creation of clinically useful diabetes risk assessment tools by combining Brier scores for calibration and DeLong/McNemar tests for statistical significance. Below is a comparison of different ways to predict the risk of diabetes (see Table 1).

Table 1 Comparison of Recent Diabetes Classification Studies and the Present Study on diabetes risk prediction with the proposed leakage-free machine learning framework.

Study	Datasets	Models Used	Evaluation Strategy	Calibration Analysis	Statistical Test	Best Model
[10]	DS 2[2]	Ensemble	Split (80/20)	Not Reported	Yes	Acc:98.2%
[16]	DS 1[1]	RNN	Split (70/30)	Not Reported	Not Reported	Acc:89.2%
[13]	DS 1	Random Forest	10-Fold CV	Not Reported	Not Reported	Acc:79.1%
[5]	DS 3 [3]	SVM	Split (80/20)	Not Reported	Not Reported	Acc:94.8%
[15]	DS 1	Random Forest	10-fold CV	Not Reported	Not Reported	Acc:95.3%
[14]	40 Papers	Various ML	Mixed	Rare ($\leq$2)	Not Reported	Varied
Current	DS 1/DS 2	LR,GRNN,	Leakage-Free [4]	Yes	Yes	Acc:99.8%

[1]Pima Indians Diabetes

[2]Early-Stage Diabetes Risk

[3]Standard Diabetes Dataset

[4]Leakage-Free Multi-Seed Stratified 10-Fold CV

The comparison in Table 1 clearly demonstrates the "reliability gap" in current diabetes prediction research that this paper seeks to address. Many of the recent studies, like those by Laila et al., Reddy et al., and Ramesh et al., report impressive performance metrics. However, they often use standard train/test splits or basic cross-validation strategies that are prone to data leakage and optimistic bias.

The table makes it clear that clinical calibration, that is essential for healthcare purposes, is almost never mentioned in recent literature. Nazirun et al. discovered that

it was used in less than 2% of the papers that they examined. Conversely, the current study stands out because it uses a leakage-free evaluation framework that goes well beyond simple accuracy metrics by integrating calibration analysis and formal statistical significance testing. This ensures that the GRNN model's outstanding performance on Dataset 2 (AUC: 0.998) is based on a stringent, consistent, and clinically validated method, which creates a higher standard for transparency in the field.

2 Methods

This section discusses how the dataset was prepared, the ways it was preprocessed, exactly how the model architectures were set up, the way the validation framework was set up, precisely how the performance was measured, and the ways the statistical comparison was performed, in addition to how reproducibility procedures were set up to validate diabetes classification models in a design that didn't allow any leaks. We implemented a strict validation protocol that prevented all information from testing subsets from affecting the training process. This was valid for all preprocessing and model evaluation steps.

2.1 Dataset Description

The current study used two publicly accessible diabetes datasets to evaluate the proposed framework. The initial source is the Pima Indians Diabetes Database [17], which contains 768 records from women who are 21 years old or older and live near Phoenix, Arizona. The dataset includes physiological measurements like glucose concentration, blood pressure, insulin level, body mass index (BMI), skin thickness, diabetes pedigree function, age, and number of pregnancies. It additionally includes a binary outcome variable that shows whether or not diabetes is present. This dataset is composed of standard clinical measurements performed in a lab and is often used to compare different diabetes classification models.

The second dataset is the Early-Stage Diabetes Risk Dataset [18], which contains 520 records of individuals who were checked for early signs of diabetes. This dataset is different from the PIMA dataset because it focuses on symptom-based indicators and demographic information, such as polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, age, and gender. The outcome parameter shows whether a person has diabetes or not, which allows for a binary classification of early diabetes risk.

A summary of both datasets is provided in Table 2.

Table 2 Summary of datasets used in the study

Dataset	Samples	Features	Type of Variables	Target
DS 1[1]	786	8 predictors	Clinical measurements	Diabetes (0/1)
DS 2[2]	520	16 predictors	Symptoms + demographic	Diabetes (0/1)

[1]Pima Indians Diabetes

[2]Early-Stage Diabetes Risk

2.2 Dataset Pre-Processing

Since two datasets with different types of characterized were used, pre-processing was done separately for each dataset, while maintaining a consistent, leakage-free evaluation framework.

2.3 Pre-Processing of the PIMA Indians Diabetes Dataset

The PIMA Indians dataset contains eight continuous clinical variables and one binary target variable, indicating whether an individual has diabetes or not. Before training the model, the following steps were taken to prepare the data:

1. The values that were missing or did not make sense from a biological point of view (like glucose, BMI, or blood pressure) were treated as missing.
2. To avoid data leakage, these values were handled using the right imputation procedures based on the training data in each cross-validation fold.
3. We used min-max normalization to scale all of the numeric variables. The scaling parameters were only estimated from the training data of each fold and then used on the test fold that went with it.

This fold-wise preprocessing ensured that no information from the test data influenced model training.

2.4 Pre-Processing of the Early-Stage Diabetes Risk Dataset

The Early-Stage Diabetes Risk dataset mainly has demographic and categorical symptom indicators that are useful for finding diabetes early. The following preprocessing steps were applied:

1. Binary symptom indicators were encoded using numerical representations (0/1).
2. Categorical variables, such as gender, were encoded using appropriate schemes.
3. Continuous variables such as age were normalized using min–max scaling.

As with Dataset 1, all preprocessing steps were done in each cross-validation fold. This made sure that the test data was never seen during training.

2.5 Hyperparameter Configuration

Model Type	Hyperparameter	Configuration	Selection Rationale
Logistic Regression	Regularization	L2	Stability and preventing overfitting.
RF	Number of Trees	100	Balanced computational cost and performance.
FFNN / BPNN	Hidden Layer Size	Predefined input	Optimized for input feature space.
FFNN / BPNN	Activation Function	Sigmoid	Standard for binary classification tasks.
GRNN	Spread Parameter	Empirically Selected	Maximizing local generalization.
ABC-FFNN	Population Size	Predefined	Based on ABC optimization standards.
ABC-FFNN	Iteration Count	Fixed across runs	Ensuring reproducibility and convergence.

Table 3 Hyperparameter Configuration for Evaluated Models

All hyperparameters remained constant across folds to ensure consistent evaluation.

2.6 Leakage-Free Evaluation Protocol

In order to make sure that the model evaluation was accurate, all of the preprocessing steps were done after the training–testing split in each cross-validation fold. This protocol makes sure that no information from the test folds affects preprocessing, training the model, or tuning the hyper-parameters.

Specifically:

1. The dataset was first partitioned using stratified K-fold cross-validation.
2. Preprocessing parameters were learned only from the training subset.
3. These parameters were then applied to the corresponding test subset.
4. Model training and evaluation were performed independently for each fold.

This procedure prevents information leakage and ensures an unbiased estimate of model performance.

Algorithm 1 Leakage-Free Preprocessing and Model Evaluation

1: **Input:** Dataset $D = (\mathrm{X}, y)$, number of folds K
2: **Output** Cross-validated model performance metrics

1. Partition the dataset D into K stratified folds.
2. For each fold k=1,. . . ,K:
   - Define training set The training set is denoted as D_{train} and test set D_{test}
   - Estimate normalization parameters from training data:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

   where $X_{\min}$ and $X_{\max}$ are computed only from D_{train}
3. Apply the learned transformation to both:
   - D_{train}
   - D_{test}
4. Train the machine learning model using normalized D_{train}
5. Evaluate the model using unseen D_{test}
6. Store performance metrics (Accuracy, Sensitivity, Specificity, AUC, F1-score)
7. Repeat across all folds and random seeds
8. Report mean ± standard deviation across runs

This evaluation strategy prevents information leakage between training and testing data and ensures reliable estimation of model generalization performance [19].

The leakage-free preprocessing and evaluation procedure adopted in this study is illustrated in Figure 1, which contrasts the typical machine learning evaluation pipeline with the fold-wise preprocessing strategy used in this work.

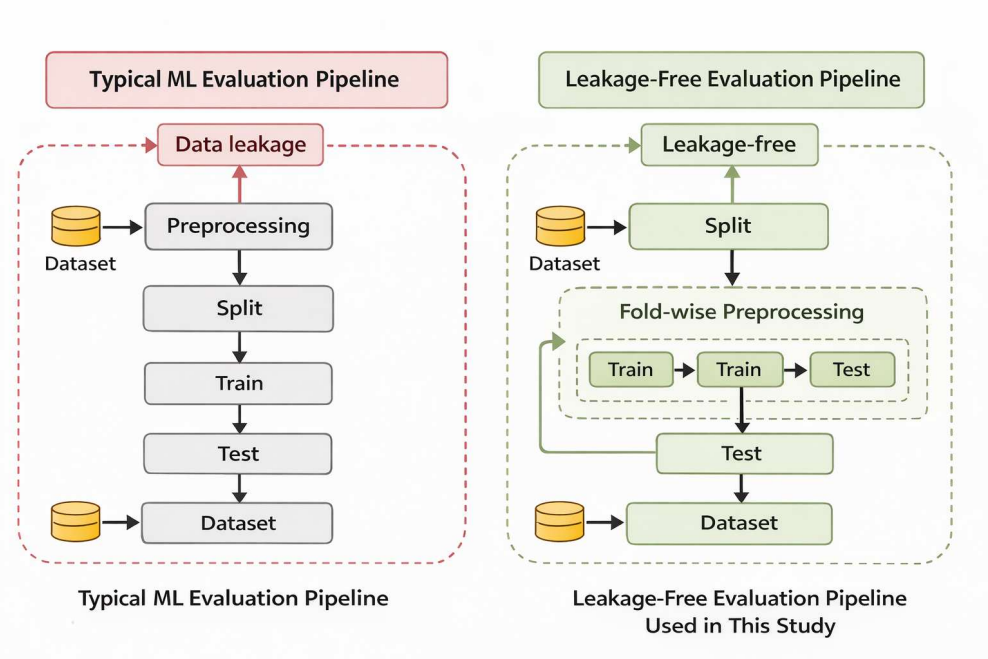


Fig. 1 Illustrates the difference between a typical machine learning evaluation pipeline and the leakage-free evaluation framework adopted in this study.

2.7 Machine Learning Approaches

2.7.1 Back Propagation Neural Network

A lot of neural network models use a multi-layer feedforward network. BP usually takes place on the output layer. The weights and thresholds are changed by sending errors from the output neurons back to the input layer and using them to change the weights of the input layer. This gives you the errors of the hidden layer and the least error sum of squares [20].

2.7.2 Generalized Regression Neural Network

GRNNs, however, don't need to be trained again and again like back-propagation networks do. It determines a random function over the input-output pairs and employs the training data to determine what the function represents. Additionally, as the size of the training set grows, the error goes down, leaving only small functional limits. There are usually four layers in a GRNN: input, pattern, summation, and output [21].

2.7.3 Feed Forward Neural Network

A conventional neural network is based on machine learning technologies that work well and can learn from a lot of data. Feed-forward neural networks take data from the input layer and move it through the network's hidden layers until it gets to the

output layer. Neurons that hold two values, a bias and weights, transmit data values across a network from one layer to the other. The neurons interconnect the nodes in a layer to the next layer. Additionally, there are no connections between the (n+1) layer and the (n) layer in a Feed Forward Neural Network. In this case, there is no feedback between the layers, and data is stored and mapped in the forward direction without returning to the previous layer's data status [22].

2.7.4 Logistic Regression

Logistic Regression (LR) is a statistical classification model that is frequently employed in medical prediction tasks, as it is simple to comprehend and can predict the probability of a binary outcome. The model employs a logistic sigmoid function to turn the log-odds of the target variable into probabilities ranging from 0 to 1. It does this by linearly combining the input features. Logistic Regression is widely utilized in clinical risk prediction models due to its provision of clear decision boundaries and interpretable feature contributions. In this study, we used Logistic Regression as a baseline classifier to set a simple but useful standard for judging how well neural network models can predict outcomes in the same cross-validation framework without any leaks. The researchers utilized similar methods in studies that use clinical datasets to predict diabetes using machine learning techniques [23].

2.7.5 Random Forest

Random Forest (RF) is an ensemble learning method that builds several decision trees during training and incorporates their predictions to make classification more accurate and stable. Each tree is trained on a bootstrap sample of the dataset, as well as when the nodes are split, random feature subsets are used. This results in trees less correlated and allows them to generalise better. Random Forest models can find nonlinear relationships and complex interactions between features that are often found in clinical datasets. Random Forest algorithms have been widely used in healthcare, including predicting diabetes risk, because they are exceptionally effective at making accurate predictions and tend to overfit. This study utilised Random Forest as a nonlinear baseline model for comparison against neural network architectures within the identical, leakage-free evaluation framework [24].

2.7.6 Cross-Validation Strategy

A multi-seed stratified 10-fold cross-validation strategy was employed for both datasets to obtain reliable and unbiased estimations of how effectively a model performed. The application of machine learning studies typically suggest employing stratified cross-validation to obtain reliable outcomes when testing models [25].

Stratified partitioning: The original class distribution remained unchanged while splitting each dataset into 10 folds.

Training and testing: In each iteration, 9 folds were used for training, and 1 fold was used for testing, ensuring that every sample served as test data exactly once.

Multi-seed evaluation: The cross-validation process was carried out additionally with multiple random seeds in order to make the evaluation more reliable.

- 20 seeds: Logistic Regression, Random Forest, Plain FFNN, GRNN.
- 5 seeds: BPNN, ABC-FFNN.

This resulted in:

- 20 × 10-fold = 200 model evaluations for baseline models and FFNN/GRNN.
- 5 × 10-fold = 50 model evaluations for BPNN and ABC-FFNN.

In addition, the model evaluation details are illustrated in Table 4.

2.8 Leakage-free Preprocessing and Evaluation Algorithm

All preprocessing steps (median imputation and min–max normalization) were fitted only on the training folds and then applied to the corresponding test fold. The complete evaluation process, including the algorithm, is summarized below:

Algorithm 2 Leakage-Free Model Evaluation

1: **Input:** Dataset $D = (\mathrm{X}, y)$, number of folds K

1. Partition the dataset into stratified folds
2. For each fold:
   - Define training and testing sets
   - Estimate preprocessing parameters from training data
   - Apply transformations to testing data
   - Train model
   - Evaluate predictions
3. Store performance metrics
4. Repeat across seeds
5. Report mean ± standard deviation

This algorithm ensures complete separation between training and testing data.

Consistency across datasets: The same evaluation protocol was applied to both Dataset 1 (Pima Indians Diabetes Dataset) and Dataset 2 (Early-Stage Diabetes Risk Dataset) to ensure a fair comparison despite differences in data characteristics.

No test data were used during tuning. All models were executed under multiple independent random seeds, and performance is reported as:

$$M = \bar{m} \pm \sigma \tag{2}$$

Where:

- $\overline{m}$ represents the mean performance metric across seeds
- σ represents the standard deviation across seeds

2.9 Statistical Analysis

Statistical comparisons were conducted using:

- DeLong Test — AUC comparison
- McNemar Test — classification accuracy
- Paired t-test
- Wilcoxon signed-rank test

Statistical significance was defined as:

$$p < 0.05 \tag{3}$$

These tests were used to verify whether observed performance differences were statistically meaningful.

2.10 Leakage-Free Diabetes Prediction Workflow

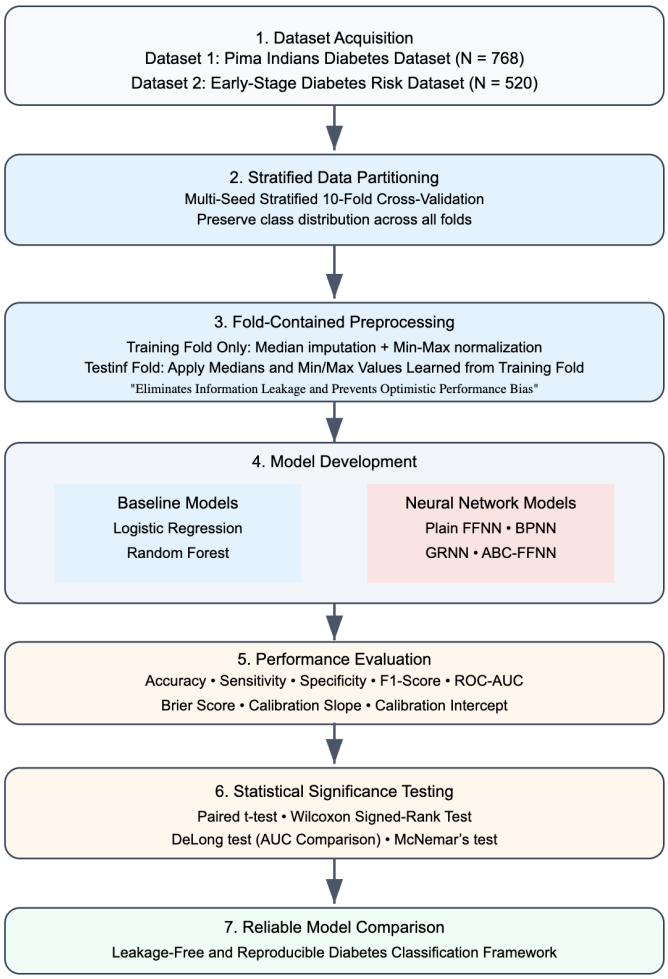


Fig. 2 Workflow of the proposed leakage-free diabetes classification framework. Multi-seed stratified 10-fold cross-validation was used to split the datasets into groups. To keep information from leaking, all preprocessing steps were done only on training folds. We used classification, discrimination, and calibration metrics to evaluate the models, and then we used statistical testing to make sure that the comparisons between models were reliable.

2.11 Model Evaluation

Evaluation of both classical and neural network models using an identical leakage-free preprocessing and cross-validation framework facilitates an equitable comparison of their predictive performance. It provides an overall view of how effectively the model performs in predicting diabetes.

Table 4 Model Architectures and Configuration Parameters for Both Datasets

Model	Datasets	Arch (I–H–O)	Training Algorithm	Seeds	Key Parameters
LR	Dataset 1[1]	16-1	LBFGS	20	L2 reg, max iter = 1000
LR	Dataset 2[2]	8-1	LBFGS	20	L2 reg, max iter = 1000
RF	Dataset 1	8→T→1	Bagging	20	Trees=200, Min leaf=5, all predictors
RF	Dataset 2	16→T→1	Bagging	20	Trees=200, Min leaf=5
FFNN	Dataset 1	8–200–1	SCG	20	Cross-entropy, 85/15 split
FFNN	Dataset 2	16–200–1	SCG	20	Cross-entropy, 85/15 split
BPNN	Dataset 1	8–200–1	LM	5	Mean squared error loss
BPNN	Dataset 2	16–200–1	LM	5	Mean squared error loss
GRNN	Dataset 1	8–200–1	RBF	20	Spread parameter = 0.2
GRNN	Dataset 2	16–200–1	RBF	20	Spread parameter = 0.2
ABC-FFNN	Dataset 1	8–200–1	ABC + FFNN	5	ABC for initial weights
ABC-FFNN	Dataset 2	16–200–1	ABC + FFNN	20	Hidden neurons = 200

[1]Pima Indians Diabetes

[2]Early-Stage Diabetes Risk

Table 4 demonstrates that baseline models such as logistic regression and random forests do not employ hidden layers. This implies that their architectures are only based on the total number of input features and the output prediction. Neural network models had one hidden layer with 200 neurons. This indicates the aforementioned PIMA dataset had architectures of 8–200–1 and the Early-Stage Diabetes Risk dataset had architectures of 16–200–1. The size of the hidden layer was chosen based on early tests and the most common configurations used in neural network-based medical prediction studies.

To evaluate model stability and mitigate the influence of random initialization, experiments were conducted using various random seeds. 20 seeds were used to test baseline models and feed-forward neural networks. 5 seeds were used to test more complex models like BPNN and ABC-FFNN.

2.12 Performance Metrics

Performance was evaluated using accuracy, sensitivity, specificity, F1-score, ROC-AUC, Brier score, calibration intercept, and Calibration slope. Uncertainty was quantified using 2000 bootstrap resamples of pooled out-of-fold predictions. [26].

These metrics were computed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6}$$

$$F_1 = \frac{TP}{TP + FP + FN} \tag{7}$$

$$BS = \frac{1}{n} \sum_{i=1}^{n} (p_i - y_i)^2 \tag{8}$$

This is the most common form, where n is the number of predictions, p_i is the predicted probability, and y_i is the actual outcome (0 or 1)

Calibration intercept and slope were estimated using logistic recalibration: α β

$$(\text{logit}(P(y_i = 1|\pi_i)) = \alpha + \beta \cdot logit(\pi_i) \tag{9}$$

Performance metrics were calculated for each seed and fold combination. The final reported results correspond to the mean and standard deviation across seeds, expressed as $\text{Metric}_{\text{mean}} \pm \text{SD}$.

In addition to these classification metrics, Receiver Operating Characteristic (ROC) analysis was conducted, and the Area Under the ROC Curve (AUC) was computed to evaluate the models' ability to discriminate between diabetic and non-diabetic cases across different decision thresholds.

The same evaluation metrics and reporting procedure were applied to both datasets to ensure a consistent and fair comparison of model performance.

3 Result

3.1 Experimental Evaluation Overview

This section presents the results of the tests that were done to see how well the proposed machine learning framework for diabetes risk prediction works without any leaks. The analysis compares baseline machine learning models with neural network-based methodologies using two datasets: the Pima Indians Diabetes dataset (Dataset 1) and the Early-Stage Diabetes Risk dataset (Dataset 2). A leakage-free multi-seed cross-validation framework was used to test the model's performance using several discrimination, classification, and calibration metrics. We examined a range of various metrics to see how well the model worked, such as accuracy, sensitivity, specificity, F1-score, ROC–AUC, Brier score, and calibration statistics. The subsequent sections show the results in terms of how effectively the models can identify differences between the two groups, how effectively the models can calibrate, precisely how well they can compare models using multiple metrics, and how accurately they can test for statistical significance between models that are competing with each other.

3.2 Model Discrimination Performance

The receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) were used to see how well the models could tell the difference between things. These metrics gauge the models' capacity to differentiate between diabetic and non-diabetic cases at various classification thresholds.

Figure 3 shows the ROC curves for the models that did the best on both datasets. The models for Dataset 1 (PIMA) show moderate discrimination ability, with AUC values between about 0.76 and 0.84. The mean AUC value for Logistic Regression was the highest (0.832), followed closely by GRNN (0.831) and Random Forest (0.829). This shows that the top models don't perform very differently from each other.

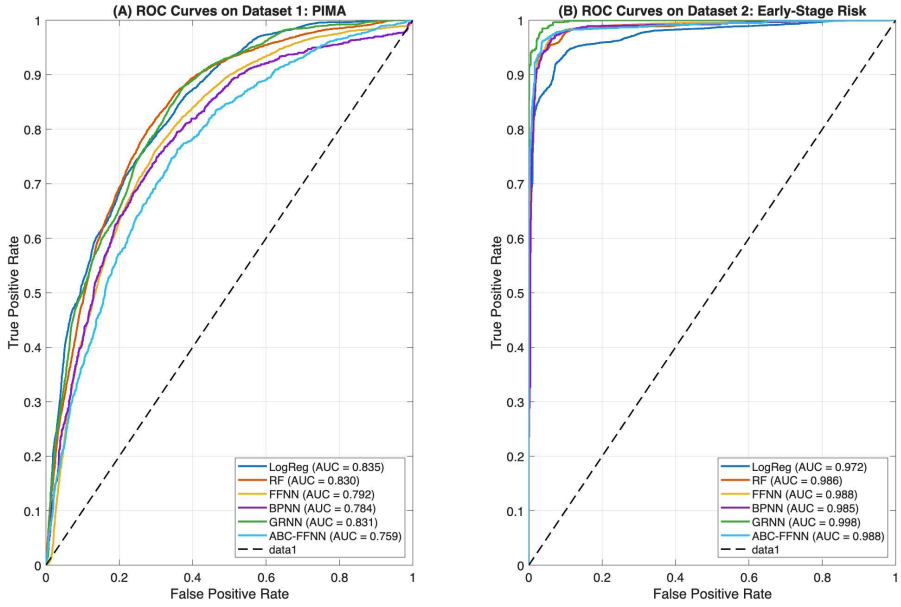


Fig. 3 ROC Curves for the top-performing models on Dataset 1 and Dataset 2 are presented. All evaluated models achieved AUC values exceeding 0.97.

In contrast, Dataset 2 showed much better discrimination performance, with all of the models tested obtaining AUC values higher than 0.97. The AUC value for GRNN was the highest (0.9976), followed closely by the AUC values for ABC-FFNN and Random Forest. Statistical comparisons, however, showed that not all pairwise differences between models were statistically significant (Table 5). This means that the differences in performance that were seen should be taken with a grain of salt.

To further illustrate cross-dataset differences, Figure 4 summarizes mean ROC–AUC values across models. In the two datasets, Dataset 2 consistently showed better discrimination performance than Dataset 1. This difference probably indicates that the symptom-based dataset has better class segregation than the clinical-measurement dataset.

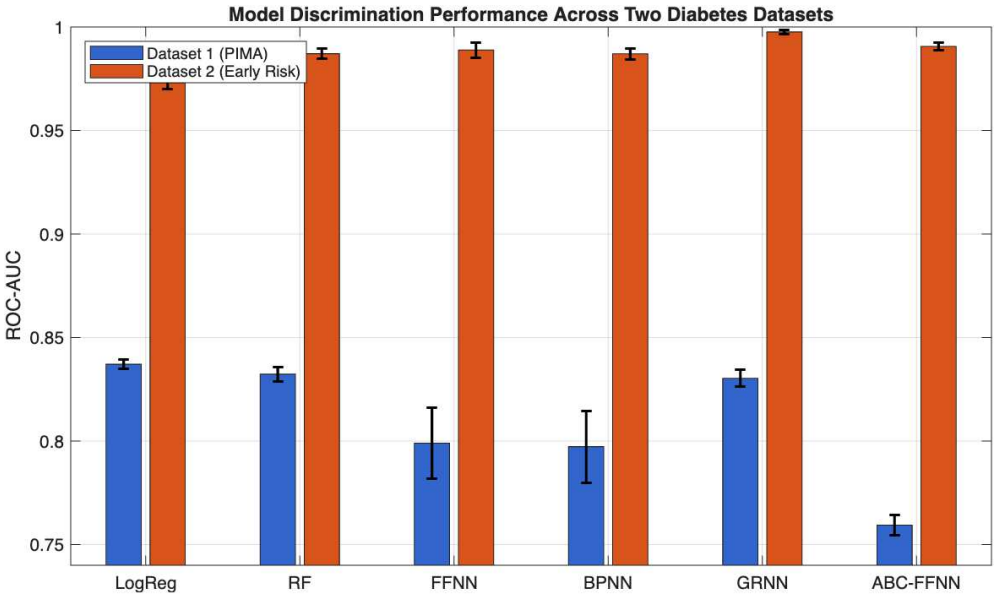


Fig. 4 Model Discrimination Performance across Two Diabetes Datasets, where Dataset 2 shows outstanding performance compared to Dataset 1 on each model.

Overall, these results indicate that several models achieved comparable discrimination performance on Dataset 1, while GRNN demonstrated the highest numerical discrimination performance on Dataset 2.

3.3 Calibration Performance

Calibration analysis was conducted to evaluate the reliability of predicted probabilities. Calibration curves and reliability diagrams are presented in Figure 5, while quantitative calibration statistics were assessed using Brier score, calibration slope, and intercept metric.

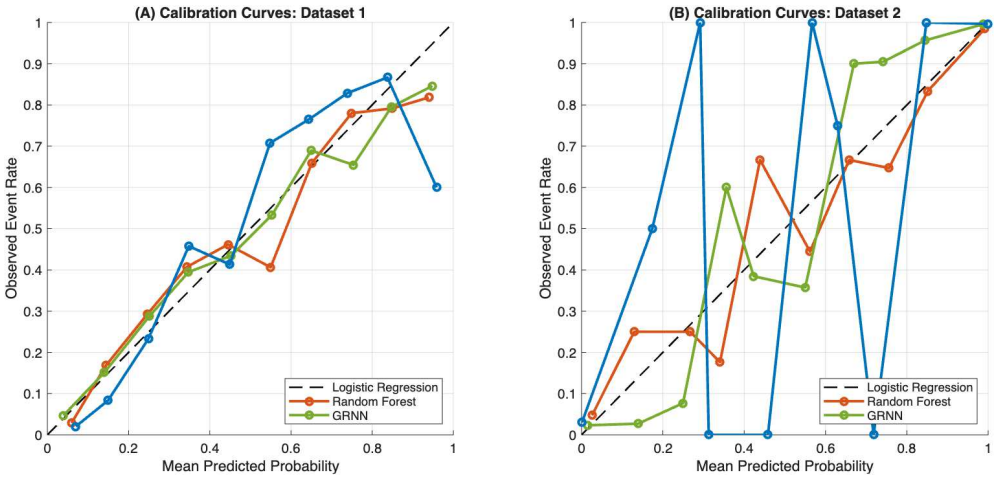


Fig. 5 Calibration Curves and Predicted Probability Distributions. The calibration curves show that high discrimination does not necessarily imply perfect probability calibration. While several models achieved strong AUC values, deviations from the diagonal reference line indicate residual miscalibration, particularly for some neural network models.

For Dataset 1, logistic regression demonstrated relatively stable calibration behaviour with predicted probabilities closely aligned the ideal calibration line across several probability bins. The GRNN model achieved the best Brier score (0.157 ± 0.001) and a calibration slope of 0.936 ± 0.007, indicating well-calibrated probability estimates.

For Dataset 2, the calibration curves indicate that most models produce highly confident predictions due to the strong separability of the dataset. The predicted probability distributions show clear clustering near the extremes of the probability range, reflecting the high classification performance observed in this dataset.

Overall, calibration analysis shows that the GRNN model gives accurate probability estimates in both cases.

3.4 Multi-Metric Model Comparison

To provide a comprehensive comparison of model performance across multiple evaluation metrics, a radar chart visualization was generated for the top-performing models. This visualization simultaneously presents six performance metrics: accuracy, sensitivity, specificity, F1-score, ROC-AUC, and calibration performance. The radar chart shown in Figure 6 compares the performance of four representative models: logistic regression, random forest, GRNN, and ABC-FFNN, across both datasets.

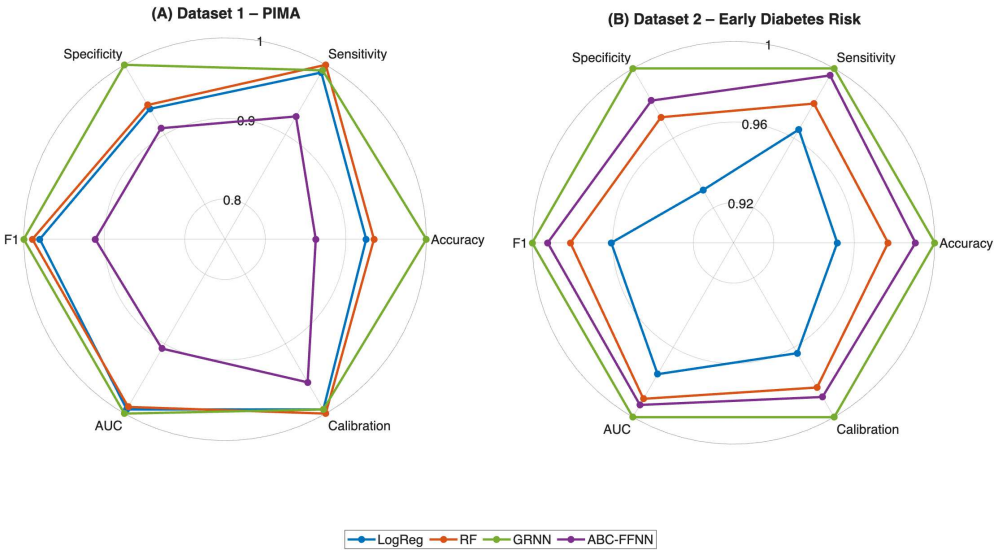


Fig. 6 Presents a radar chart comparing the top-performing models across six evaluation metrics for both datasets. The metrics include accuracy, sensitivity, specificity, F1-score, ROC–AUC, and calibration performance.

For Dataset 1, the GRNN model exhibits the most balanced performance across all metrics, particularly in sensitivity and F1-score, which are important indicators for identifying positive diabetes cases. Logistic regression and random forest demonstrate comparable performance, while the ABC-FFNN model shows slightly lower performance across several metrics.

For Dataset 2, the radar chart indicates that all models achieve consistently high performance across all metrics, reflecting the strong predictive signal present in the dataset. However, GRNN again demonstrates the most consistent performance across the evaluation dimensions.

These results highlight the robustness of the GRNN model and confirm its ability to achieve strong predictive performance across both clinical and symptom-based datasets.

3.5 Statistical Significance of Model Comparisons

To determine whether observed performance differences between models were statistically significant, several paired statistical tests were conducted.

Pairwise statistical comparisons were conducted using DeLong's test for ROC–AUC differences and McNemar's test for classification accuracy differences. Pairwise significance matrices for Dataset 1 and Dataset 2 are presented in Figures 8 and 9. These heatmaps visualize pairwise p-values obtained from DeLong tests comparing the ROC–AUC scores of different models, while detailed comparison statistics are summarized in Tables 5 and 6.

The heatmap in Figure 7 illustrates pairwise comparisons of ROC–AUC values between all evaluated models. Lower p-values (warmer colours) indicate statistically significant differences between model discrimination performances, while higher p-values (cooler colours) indicate comparable performance. For Dataset 1, most pairwise comparisons yielded moderate to high p-values, indicating relatively similar discrimination capability among several models. However, statistically significant differences

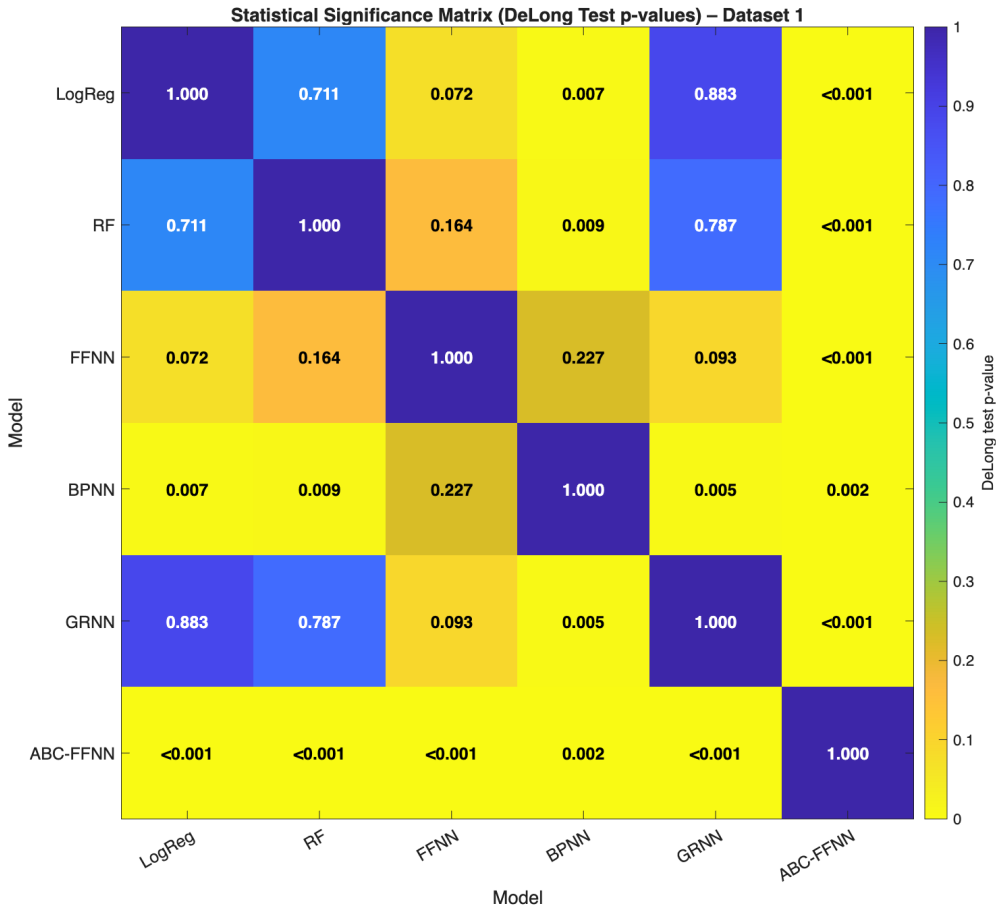


Fig. 7 Statistical Significance Matrix (DeLong Test p-values) on Dataset 1 (PIMA)

$p < 0.05$ were observed between selected neural network models and baseline classifiers, suggesting localized performance improvements rather than uniformly superior performance.

Table 5 Statistical comparison of models on Dataset 1 (PIMA)

Model Comparison	ΔAUC	DeLong z	p-value	Acc. Diff	McNemar p-value	Interpretation
LR vs RF	−0.003	0.370	0.711	0.003	0.916	Not significant
LR vs FFNN	−0.018	1.800	0.072	−0.020	0.159	Not significant
LR vs BPNN	−0.032	2.683	0.007	−0.020	0.164	LR better
LR vs GRNN	−0.001	0.147	0.883	−0.005	0.689	Not significant
LR vs ABC-FFNN	−0.067	4.572	< 0.001	−0.050	0.003	LR better
RF vs FFNN	−0.015	1.393	0.164	−0.022	0.143	Not significant
RF vs BPNN	−0.028	2.615	0.009	−0.022	0.115	RF better
RF vs GRNN	0.002	0.271	0.787	0.003	0.921	Not significant
RF vs ABC-FFNN	−0.064	4.899	< 0.001	−0.052	0.001	RF better
FFNN vs BPNN	−0.014	1.209	0.227	0.000	0.921	Not significant
FFNN vs GRNN	0.017	1.680	0.093	0.025	0.079	Not significant
FFNN vs ABC-FFNN	−0.049	3.416	< 0.001	−0.030	0.064	FFNN better
BPNN vs GRNN	0.031	2.821	0.005	0.025	0.088	GRNN better
BPNN vs ABC-FFNN	−0.035	3.093	0.002	−0.030	0.047	BPNN better
GRNN vs ABC-FFNN	−0.066	4.911	< 0.001	−0.055	0.001	GRNN better

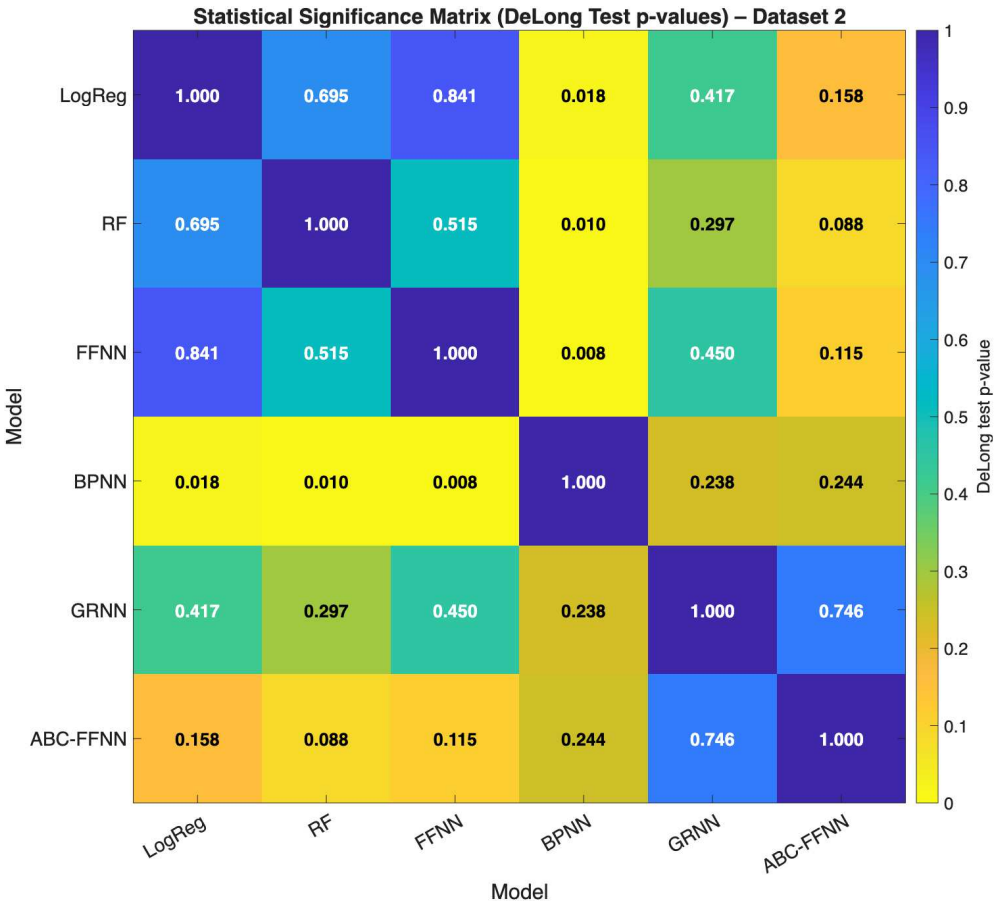


Fig. 8 Statistical Significance Matrix on Dataset 2 (Early Diabetes Risk)

The heatmap in Figure 8 visualizes pairwise statistical comparisons of ROC–AUC performance across all evaluated models. Lower p-values indicate statistically significant differences in discrimination ability, while higher values indicate comparable model behaviour. In contrast to Dataset 1, Dataset 2 exhibits fewer statistically significant differences across many model pairs, reflecting the consistently high predictive performance achieved across models. Nevertheless, statistically significant improvements were observed for selected neural network architectures, particularly in comparisons involving the GRNN model, highlighting its superior discrimination capability in the symptom-based dataset.

Table 6 Statistical comparison of models on Dataset 2 (Early Diabetes Risk)

Model Comparison	ΔAUC	DeLong z	p-value	Acc. Diff	McNemar p-value	Interpretation
LR vs RF	−0.005	0.392	0.695	0.015	0.216	Not significant
LR vs FFNN	0.002	0.201	0.841	0.023	0.045	FFNN better
LR vs BPNN	0.031	2.363	0.018	0.023	0.052	BPNN better
LR vs GRNN	0.013	0.811	0.417	0.019	0.144	Not significant
LR vs ABC-FFNN	0.018	1.413	0.158	0.023	0.082	Not significant
RF vs FFNN	0.007	0.650	0.516	0.008	0.455	Not significant
RF vs BPNN	0.036	2.577	0.010	0.008	0.524	BPNN better
RF vs GRNN	0.018	1.042	0.297	0.004	0.845	Not significant
RF vs ABC-FFNN	0.023	1.707	0.088	0.008	0.524	Not significant
FFNN vs BPNN	0.029	2.638	0.008	0.000	1.000	BPNN better
FFNN vs GRNN	0.011	0.755	0.450	−0.004	0.791	Not significant
FFNN vs ABC-FFNN	0.016	1.578	0.115	0.000	1.000	Not significant
BPNN vs GRNN	−0.018	1.180	0.238	−0.004	0.804	Not significant
BPNN vs ABC-FFNN	−0.013	1.164	0.244	0.000	1.000	Not significant
GRNN vs ABC-FFNN	0.005	0.324	0.746	0.004	0.815	Not significant

3.6 Confusion Matrix Analysis

To further examine class-wise diagnostic behavior, pooled confusion matrices were constructed across all outer test folds and random seeds under the leakage-free evaluation framework. Figure 9 illustrates the row-normalized confusion matrices for all evaluated models, highlighting class-specific prediction behavior.

For Dataset 1, Logistic Regression demonstrated strong identification of non-diabetic cases 87.9%, while Random Forest achieved the highest detection of diabetic cases 63.0%. GRNN achieved the highest specificity 91.4% but comparatively lower sensitivity 48.0%, indicating strong identification of non-diabetic individuals but reduced detection of positive cases.

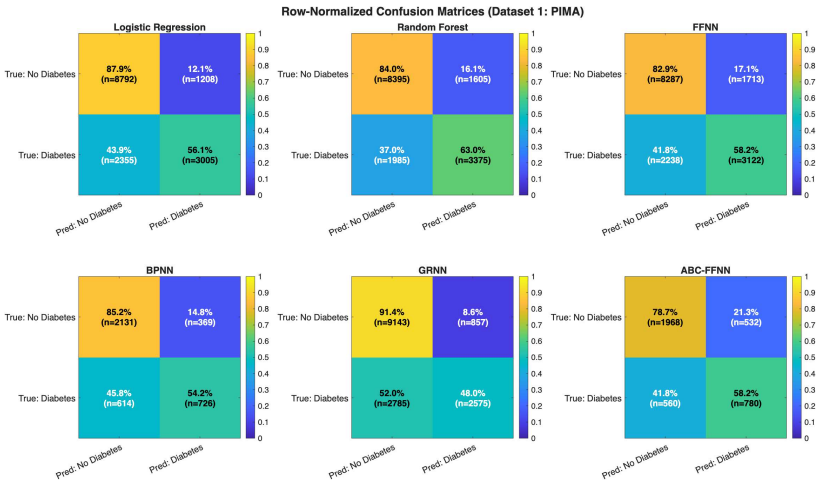


Fig. 9 Row-normalized confusion matrices for baseline and neural network models under leakage-free multi-seed cross-validation. Percentages represent class-wise prediction proportions, and counts in parentheses indicate pooled observations aggregated across folds and random seeds.

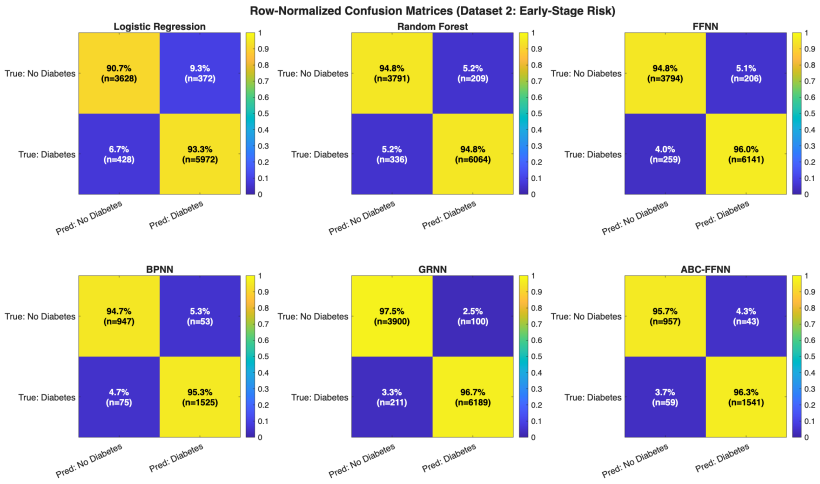


Fig. 10 Row-normalized confusion matrices for baseline and neural network models on the Early-Stage Diabetes Risk dataset under leakage-free multi-seed cross-validation. Each cell represents the proportion of predictions for the corresponding true class.

For Dataset 2, all evaluated models demonstrated high sensitivity and specificity values. GRNN and ABC-FFNN achieved balanced classification performance, with sensitivity and specificity values exceeding 0.95. These results are consistent with the strong discrimination performance observed in Dataset 2.

Overall, confusion matrix analysis highlights meaningful differences in class-wise diagnostic behavior and emphasizes the importance of considering sensitivity–specificity trade-offs in model evaluation.

3.7 Decision Curve Analysis

Decision curve analysis was conducted to evaluate the potential decision relevance of selected models across clinically relevant threshold probabilities. The results are presented in Figure 11.

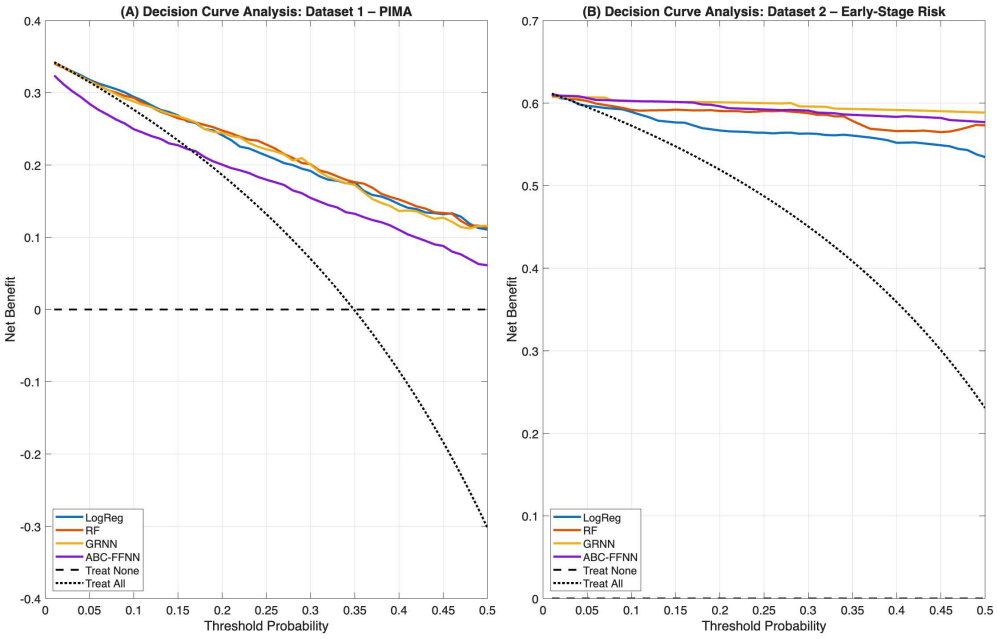


Fig. 11 Decision Curve Analysis Evaluation Across Clinically Relevant Threshold Probabilities for both Datasets.

For Dataset 1, Net Benefit values remained relatively modest across threshold ranges, reflecting the moderate discrimination performance observed in this dataset. Logistic Regression and Random Forest demonstrated slightly improved Net Benefit compared with alternative models at intermediate thresholds.

For Dataset 2, higher Net Benefit values were observed across a broader range of threshold probabilities, reflecting the strong predictive signal present in the symptom-based dataset. GRNN and ABC-FFNN demonstrated slightly higher Net Benefit across several threshold values compared with baseline models.

However, these findings should be interpreted as preliminary evidence of potential decision relevance rather than definitive clinical utility. External validation on independent datasets is required before drawing conclusions regarding real-world clinical implementation.

4 Discussion

4.1 Interpretation of Model Performance Across Datasets

This study evaluated the performance of multiple machine learning and neural network models for diabetes classification using a leakage-free multi-seed cross-validation framework across two publicly available datasets. The results demonstrate that several models achieved comparable predictive performance on Dataset 1 (Pima Indians Diabetes dataset), while consistently higher discrimination performance was observed on Dataset 2 (Early-Stage Diabetes Risk dataset).

For Dataset 1, performance differences among Logistic Regression, Random Forest, and GRNN were relatively small. Logistic Regression achieved the highest mean ROC–AUC value, while GRNN demonstrated competitive classification performance, particularly in terms of specificity and Brier score. However, statistical testing indicated that the discrimination difference between Logistic Regression and GRNN was not statistically significant ($p = 0.883$), suggesting that these models exhibited comparable predictive capability. These findings emphasize that classical baseline models remain competitive when evaluated under leakage-free conditions and robust validation schemes.

In contrast, substantially higher performance levels were observed across all models for Dataset 2. GRNN achieved the highest numerical ROC–AUC and strong calibration performance, although multiple pairwise comparisons did not demonstrate statistically significant differences between leading models. These results suggest that while GRNN demonstrated favorable performance characteristics, superiority claims should be interpreted within the broader context of dataset characteristics and statistical uncertainty.

4.2 Influence of Dataset Characteristics on Model Performance

Comparing Dataset 1 and Dataset 2 gives us significant information about how the structure of data affects how well it can predict outcomes. Dataset 1 mainly consists of clinical measurements, such as physiological and biochemical variables, that might have distributions that are similar for both diabetic and non-diabetic people. This natural overlap is probably what makes the models' discrimination performance just moderate.

Dataset 2, on the other hand, has indicators based on symptoms,including polyuria, polydipsia, abrupt weight loss, and weakness. These characteristics are closely related to diabetes status, which means that classes can be easily separated. Because of this, almost all the models did a great job in separating Dataset 2, with ROC–AUC values above 0.97. These results show that there is a strong predictive signal in the dataset,

but they also show that the distance between features and targets may have an effect, where highly predictive symptoms might make classification tasks easier.

It is important to note that near-perfect performance in dataset 2 does not mean that the model can predict anything. Instead, these results show how important it is to test prediction models on different datasets that have different demographic details and distinctive feature complexity.

4.3 Calibration and Reliability of Probability Estimates

Beyond discrimination performance, calibration analysis demonstrated that several models produced reliable probability estimates across both datasets. Logistic Regression demonstrated stable calibration behavior, consistent with its well-established probabilistic formulation. GRNN achieved favorable Brier score values and calibration slope estimates, indicating reliable probability predictions in both datasets.

These findings underscore the importance of evaluating calibration alongside discrimination metrics, particularly in medical decision-making contexts where predicted probabilities inform clinical judgment. Models that demonstrate strong discrimination but poor calibration may still produce unreliable risk estimates, highlighting the need for comprehensive performance evaluation.

4.4 Statistical Significance of All Model Comparisons

The observed differences in performance were clarified using various statistical methods.These included DeLong's test to compare ROC–AUC values and McNemar's test to assess classification accuracy.

In Dataset 1, only a few statistically significant differences were found among the best-performing models. The results indicate that the algorithms, when evaluated using consistent validation methods that prevent data leakage, showed similar predictive abilities.

Dataset 2 showed more statistically significant differences, particularly when comparing certain neural network models.

However, comparisons between Logistic Regression and GRNN not show statistically significant differences, indicating that these models performed similarly, even though their ROC–AUC values were different.

The results highlight the importance of using statistical tests when comparing machine learning models. Relying solely on differences in numerical performance can lead to incorrect conclusions about which model is better.

4.5 Decision Curve Analysis and Potential Decision Relevance

We used decision curve analysis to see if the evaluated models could help make decisions at clinically relevant threshold probabilities. The results demonstrated that a few models had a higher net benefit than baseline strategies in certain threshold ranges.

In Dataset 2, GRNN and ABC-FFNN showed a little more net benefit across intermediate threshold values than the baseline models. However, these results should be taken with a grain of salt because decision curve analysis only provides us a rough

estimate of the potential significance of a decision, not solid proof of its clinical usefulness.

It's important to note that the datasets used in this study are retrospective benchmark datasets, not clinical cohorts that were collected in the future. Before any model can be used to help with clinical decisions in the real world, it must be tested on different populations.

4.6 Comparison with Previous Studies

The findings of this study are broadly consistent with prior research demonstrating strong predictive performance of machine learning models for diabetes classification tasks. However, many earlier studies have relied on simplified validation procedures, including single train–test splits or preprocessing applied before cross-validation, which may introduce optimistic bias.

In contrast, the present study employed a leakage-free evaluation framework in which preprocessing operations were restricted to training folds and applied independently to test data. This methodological design reduces the risk of data leakage and provides more reliable estimates of model generalization performance. Additionally, the use of multi-seed repeated cross-validation and statistical comparison testing provides a more rigorous evaluation framework compared with many prior diabetes prediction studies. These methodological differences represent an important contribution of the present work and support more transparent comparison of predictive algorithms.

4.7 Study Limitations

There are certain considerations to keep in mind when looking at the results of this study.

First, the datasets utilised in this study are publicly accessible benchmark datasets that may not comprehensively reflect real-life clinical samples. The strong predictive performance seen on Dataset 2 may be due to the inclusion of symptom-based variables that are quite effective at demonstrating if someone has diabetes or not. This particular type of feature-target proximity may make classification easier and make it harder to apply to other screening scenarios.

Second, different types of conventional models had different numbers of repeated training seeds because of limits on how much computing power they could use. This design made it easy to evaluate models, but different numbers of repetitions may affect how accurate performance estimates are across model families. Third, we chose the hyperparameter settings based on early tests and values that have been reported in other studies. While these settings generated consistent performance results, varying parameter values might produce divergent performance outcomes.

Fourth, this study was more about binary classification of diabetes status than it was about predicting future disease risk over time. So, the results should be looked at in terms of classification instead of long-term risk prediction.

Future studies should use larger, prospectively collected clinical datasets to test these models and add nested hyperparameter optimisation to make the methods even more rigorous.

4.8 Implications for Future Research and Model Evaluation

The study's results highlight how crucial it is to adopt validation methods that don't let any leaks happen in medical machine learning research. Even tiny errors in preprocessing can lead to overly optimistic performance predictions, especially when datasets have features that are highly effective at predicting.

In order to make sure the model's evaluation is accurate and reliable, future research must focus on external validation, calibration assessment, and statistical comparison of competing models. Also, adding explainability methods and clinical workflow concerns could make predictive models more useful in healthcare settings.

The results reveal that differences in model performance are often caused by the dataset. This means that we need to evaluate prediction algorithms on more than one benchmark dataset.

Table 7 Performance comparison of models on Dataset 1 (PIMA) and Dataset 2 (Early-Stage Diabetes Risk)

Model	Dataset	Accuracy	Sensitivity	Specificity	F1-score	ROC AUC	Brier Score	Cal. Slope	Cal. Int.
LR	DS 1	0.7680 ± 0.0038	0.5606 ± 0.0078	0.8792 ± 0.0033	0.6278 ± 0.0068	0.8371 ± 0.0023	0.1569 ± 0.0005	0.9364 ± 0.0074	-0.0336 ± 0.0050
LR	DS 2	0.9231 ± 0.0044	0.9331 ± 0.0048	0.9070 ± 0.0068	0.9372 ± 0.0036	0.9730 ± 0.0029	0.0605 ± 0.0019	0.7830 ± 0.0291	-0.0425 ± 0.0272
RF	DS 1	0.7663 ± 0.0054	0.6297 ± 0.0115	0.8395 ± 0.0061	0.6528 ± 0.0085	0.8323 ± 0.0036	0.1584 ± 0.0017	0.8201 ± 0.0189	-0.0808 ± 0.0130
RF	DS 2	0.9476 ± 0.0059	0.9475 ± 0.0057	0.9477 ± 0.0119	0.9570 ± 0.0048	0.9871 ± 0.0026	0.0413 ± 0.0023	1.2380 ± 0.1085	0.2613 ± 0.0660
P-FFNN	DS 1	0.7428 ± 0.0159	0.5825 ± 0.0312	0.8287 ± 0.0341	0.6124 ± 0.0156	0.7994 ± 0.0193	0.1840 ± 0.0172	0.4599 ± 0.2098	-0.2840 ± 0.1793
P-FFNN	DS 2	0.9553 ± 0.0069	0.9595 ± 0.0082	0.9485 ± 0.0106	0.9635 ± 0.0056	0.9889 ± 0.0037	0.0357 ± 0.0065	0.5393 ± 0.0803	0.1184 ± 0.1655
BPNN	DS 1	0.7440 ± 0.0116	0.5418 ± 0.0417	0.8524 ± 0.0110	0.5957 ± 0.0280	0.7973 ± 0.0173	0.1824 ± 0.0116	0.4718 ± 0.2025	-0.2149 ± 0.1205
BPNN	DS 2	0.9508 ± 0.0119	0.9531 ± 0.0193	0.9470 ± 0.0115	0.9597 ± 0.0101	0.9870 ± 0.0027	0.0385 ± 0.0074	0.5579 ± 0.0626	0.2224 ± 0.2286
GRNN	DS 1	0.7629 ± 0.0036	0.4804 ± 0.0102	0.9143 ± 0.0042	0.5857 ± 0.0081	0.8323 ± 0.0021	0.1626 ± 0.0007	1.3266 ± 0.0301	0.2724 ± 0.0163
GRNN	DS 2	0.9701 ± 0.0053	0.9670 ± 0.0082	0.9750 ± 0.0040	0.9755 ± 0.0045	0.9976 ± 0.0009	0.0247 ± 0.0034	0.3402 ± 0.0162	0.9427 ± 0.1485
ABC-FFNN	DS 1	0.7156 ± 0.0039	0.5821 ± 0.0183	0.7872 ± 0.0092	0.5881 ± 0.0093	0.7636 ± 0.0063	0.2177 ± 0.0016	0.2293 ± 0.0149	-0.3764 ± 0.0195
ABC-FFNN	DS 2	0.9608 ± 0.0044	0.9631 ± 0.0026	0.9570 ± 0.0084	0.9680 ± 0.0036	0.9906 ± 0.0018	0.0360 ± 0.0038	0.3001 ± 0.0109	0.3923 ± 0.0851

Table 7 summarizes the predictive performance of the evaluated baseline and neural network models on the two diabetes datasets under the proposed leakage-free multi-seed stratified cross-validation framework. The values are reported as mean ± standard deviation across repeated multi-seed stratified 10-fold cross-validation runs.

5 Conclusion

This study introduced a leakage-free machine learning framework for reliable diabetes classification across heterogeneous datasets. The framework integrates fold-contained pre-processing, multi-seed stratified cross-validation, calibration analysis, and statistical testing to improve evaluation reliability.

On Dataset 1 (Pima Indians Diabetes Dataset), models demonstrated moderate discrimination performance (ROC–AUC ≈ 0.76–0.84), with GRNN achieving AUC $= 0.837 \pm 0.002$, Logistic Regression AUC $= 0.832$, and Random Forest AUC $= 0.829$, with small performance differences across top models. On Dataset 2 (Early-Stage Diabetes Risk Dataset), all models demonstrated high discrimination (ROC–AUC ≥ 0.97), with the GRNN achieving an AUC of 0.998. However, several pairwise comparisons were not statistically significant, indicating comparable performance among leading models.

The high performance observed on Dataset 2 likely reflects strong feature–target relationships and may not generalize to broader clinical settings. Therefore, external validation remains essential. Overall, the primary contribution of this work is the development of a reproducible and statistically grounded evaluation framework for diabetes classification studies. Future work will focus on external validation, expanded datasets, and improved interpretability.

Supplementary information. Data Availability and Code: The MATLAB code used to implement the leakage-free ML framework for diabetes risk prediction and reproduce the experiments in this study is available at: https://github.com/emmadigital/leakage-free-diabetes-risk-prediction

Acknowledgements. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

During the preparation of this manuscript/study, the author(s) used MATLAB, R2025b, for the purposes of performing analysis and the MATLAB AI tool code assistance.

Declarations

- Funding: This research received no external funding.
- Conflict of interest: The authors declare no conflicts of interest.

- Ethics approval and consent to participate: Both Authors conducted the experiments and prepared the manuscript.
- Consent for publication: The authors have reviewed and edited the output and take full responsibility for the content of this publication.
- Data availability: Both dataset used in this study are publicly available at the UCI Machine Learning Repository: https://www.kaggle.com/uciml/pima-indians-diabetes-database, https://doi.org/10.24432/C5VG8H
- Materials availability: Not applicable
- Code availability: The code is available at https://github.com/emmadigital/leakage-free-diabetes-risk-prediction
- Author contribution: Conducted the experiments and prepared the manuscript.

References

[1] Khanal, M., Bhandari, P., Rawal, L., Pandey, G., Bhandari, M., Devkota, S., De Courten, M., De Courten, B.: Poor glycemic control, cardiovascular disease risk factors and their clustering among patients with type 2 diabetes mellitus:. A cross-sectional study from Nepal. PLoS One, **17**(7.) (2022) https://doi.org/10.1371/journal.pone.0271888

[2] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I.: Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, **15**, 104–116 (2017) https://doi.org/10.1016/j.csbj.2017.01.008

[3] Marzouk, R., Alluhaidan, A.S., El Rahman, S.A.: An analytical predictive models and secure web-based personalized diabetes monitoring system. IEEE Access. **10.** (2022) https://doi.org/10.1109/TSP.2022.3214376

[4] Aftab, S., Alanazi, S., Ahmad, M., Khan, M.A., Fatima, A.e.a.: Cloud-based diabetes decision support system using machine learning fusion. Computers, Materials and Continua, **68(1)**, 1341–1357 (2021) https://doi.org/10.32604/cmc.2021.016814

[5] Ramesh, J., Aburukba, R., Sagahyroon, A..: A remote healthcare monitoring framework for diabetes prediction using machine learning. Health Technology, 45–57 (2021) https://doi.org/10.1049/htl2.12010

[6] Alnowaiser, K.: Improving healthcare prediction of diabetic patients using KNN imputed features and tri-ensemble model. IEEE Access **12**, 16783–16793 (2024) https://doi.org/10.1109/ACCESS.2024.3359760

[7] Jiang, L., Xia, Z., Zhu, R., Gong, H., Wang, J., Li, J., Wang, L.: Diabetes risk prediction model based on community follow-up data using machine learning. Preventive Medicine Reports **35**, 16783–16793 (2023) https://doi.org/10.1016/j.pmedr.2023.102358

[8] Ahmed, A., Khan, J., Arsalan, M., Ahmed, K., Shahat, A.A., Alhalmi, A., Naaz, S.: Machine learning algorithm-based prediction of diabetes among female population using PIMA dataset. Healthcare **13**(1), 37 (2025) https://doi.org/10.3390/healthcare13010037

[9] Tasin, I., Nabil, T.U., Islam, S., Khan, R.: Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters **10**, 1–10 (2023) https://doi.org/10.1049/htl2.12039

[10] Laila, U.e., Mahboob, K., Khan, A.W., Khan, F., Taekeun, W.: An ensemble approach to predict early-stage diabetes risk using machine learning:. An Empirical Study. Sensors, **22(14)**, 5247 (2022) https://doi.org/10.3390/s22145247

[11] Nipa, N., Riyad, M.H., Walliullah, S.S., Howlader, K.C., Moni, M.A.: Clinically adaptable machine learning model to identify early appreciable features of diabetes. Intelligent Medicine. **4(1)**, 22–32 (2024) https://doi.org/10.1016/j.imed.2023.01.003

[12] Kırbaş, İ., Çifci, A.: Leveraging SHAP for interpretable diabetes prediction: A study of machine learning models on the pima indians diabetes dataset. Balkan Journal of Electrical and Computer Engineering **13**(2), 128–139 (2025) https://doi.org/10.17694/bajece.1577929

[13] Mishra, S.R., Dash, S.: Machine learning based diabetes prediction using the pima indian dataset. 2nd International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), 1–6 (2024) https://doi.org/10.1109/SCOPES64467.2024.10991027

[14] Nazirun, N.N.N., al: Prediction models for type 2 diabetes progression: A systematic review. IEEE Access **2**, 161595–161619 (2024) https://doi.org/10.1109/ACCESS.2024.3432118

[15] Massari, H.E., Sabouri, Z., Mhammedi, S., Gherabi, N.: Diabetes prediction using machine learning algorithms and ontology. Journal of ICT Standardization **10(02)**, 319–338 (2022) https://doi.org/10.13052/jicts2245-800X.10212

[16] Reddy, L.C.S., Gottipalli, M., Sravanthi, P., Rajanikanth, J., Yalamarthi, G., Gurrapu, N.: Bridging horizons in diabetes prediction: A comparative exploration of machine learning and deep learning approaches in pima indian women. 2024 2nd International Conference on Advancement in Computation and Computer Technologies (InCACCT), 386–391 (2024) https://doi.org/10.1109/InCACCT61598.2024.10550977

[17] Kaggle: Pima Indians Diabetes Dataset, 2017. https://www.kaggle.com/uciml/pima-indians-diabetes-database

[18] UCI Machine Learning Repository: Early Stage Diabetes Risk Prediction

Dataset. https://uci.edu. Accessed: 2024-05-20 (2020). https://doi.org/10.24432/C5VG8H

[19] Kaufman, S., Rosset, S., Perlich, C., Stitelman, O.: Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data. (2012) https://doi.org/10.1145/2382577.2382579

[20] Li, J., H., C.J., Shi, J.Y., Huang, F.: Brief introduction of back propagation (bp) neural network algorithm and its improvement. In Advances in computer science and information engineering: **2**, 553–558 (2012) https://doi.org/10.1007/978-3-642-30223-7_87

[21] Hannan, S.A., Manza, R.R., Ramteke, R.J.: Generalized regression neural network and radial basis function for heart disease diagnosis. International Journal of Computer Applications **7(13)**, 7–13 (2010)

[22] Svozil, D., Kvasnicka, V., Pospichal, J.: Introduction to multi-layer feed-forward neural networks. **39(1)**, 43–62 (1997) https://doi.org/10.1016/S0169-7439(97)00061-0

[23] Falvo, F.R., Cannataro, M.: Explainability techniques for artificial intelligence models in medical diagnostic. IEEE Access, International Conference on Bioinformatics and Biomedicine (BIBM) **11**, 6907–6913 (2024) https://doi.org/10.1109/BIBM62325.2024.10821826

[24] Wang, S.: Diabetes prediction using random forest in healthcare. highlights in science, engineering and technology. IEEE Access **92**, 210–217 (2024) https://doi.org/10.54097/5ndh9a05

[25] Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai **14(2)**, 1137–1145 (1995)

[26] JF, C., DA, K., DG, A., DE, B., CA, G., L, H., L, I., D, L., JB, R., Vet HC.: Bossuyt pm. stard 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open **6(11)** (2016) https://doi.org/10.1136/bmjopen-2016-012799