

# Diabetic prediction using ridgelet process neural network with noise contrastive estimation

Parashiva Murthy Basavanapura Muddumadappa<sup>1</sup>, Pruthvi Palahalli Ramachandra<sup>2</sup>,  
Manjunatha Basavaraju<sup>3</sup>, Nandeesh Hallimysore Devaraj<sup>1</sup>

<sup>1</sup>Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering,  
JSS Science and Technology University, Mysuru, India

<sup>2</sup>Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, India

<sup>3</sup>Department of Industrial and Production Engineering, Sri Jayachamarajendra College of Engineering,  
JSS Science and Technology University, Mysuru, India

## Article Info

### Article history:

Received May 21, 2025

Revised Jul 28, 2026

Accepted Aug 12, 2026

### Keywords:

Data mining

Deep learning

Diabetic prediction

Feature selection

Noise contrastive estimation

Ridgelet neural network

## ABSTRACT

Globally, diabetes is the most prevalent disease and a major public health concern, often leading to serious complications like kidney disease, vision loss, and coronary issues. Deep learning (DL) and data mining techniques have emerged as reliable and highly effective methods for predicting diabetes, offering significant support to the medical field in accurate diagnosis. However, challenges like missing data and difficulties in handling high-dimensional data significantly impact diabetes prediction, resulting in poor performance. This research proposes a ridgelet process neural network (RPNN) with noise contrastive estimation (NCE) to mitigate the effects of noise and remove irrelevant features, enhancing prediction accuracy. In the pre-processing, standard scaling is involved to ensure data categorization, and missing data is handled using the interquartile range (IQR) method to improve prediction performance. The feature selection using partial least angle regression (PLAR) is utilized to handle high-dimensional data and minimize multicollinearity among features efficiently. The RPNN using a ridgelet transformer maps the data into higher-dimensional space, enabling capture of complex features and effective handling of linear patterns. The proposed method achieves better accuracy of 99.89%, 98.63%, and 99.37% on datasets like diabetes, PIMA, and diabetes 130-US when compared with existing methods like guided artificial neural network (G-ANN).

*This is an open access article under the [CC BY-SA](#) license.*



## Corresponding Author:

Manjunatha Basavaraju

Department of Industrial and Production Engineering

Sri Jayachamarajendra College of Engineering, JSS Science and Technology University

Mysuru, Karnataka 570005, India

Email: manjub@jssstuniv.in

## 1. INTRODUCTION

Diabetes is a chronic metabolic disease categorized through blood glucose levels, often influenced by lifestyle changes, dietary habits, and genetic factors [1]. Early detection of diabetes is essential, and it is categorized into 3 main types. Type 1 diabetes, called insulin-dependent diabetes, arises when the body's immune system occurrences insulin-producing cells and stops insulin production entirely [2]. Healthy individuals reduce the risk of developing diabetes through proper diet and regular exercise, while older individuals require daily medication to control blood sugar levels and manage the disease effectively [3]. The existing approach involved several machine learning (ML) approaches to analyze data from multiple perspectives, transforming it into actionable insights. This research focuses on the application of data mining

to large datasets, identifying relevant features and utilizing ML approaches to process critical applications in diabetes prediction and management [4], [5]. Diabetes involves impaired insulin function, a hormone generated through pancreatic islets, and is a major cause of heart disease, renal failure, and excessive stress, contributing to its rising prevalence [6]. The key challenge in diabetes prediction lies in handling missing data and effectively utilizing data mining techniques to process large datasets for improved predictive performance [7]. The integration of deep learning (DL) has significantly advanced various fields, particularly diabetes prediction, enhancing diagnostic accuracy, and patient outcomes [8]. Collected data undergoes preprocessing to solve missing values and standardize features, ensuring uniform data processing [9].

This research paper considers the various datasets to process large data to resolve the high dimensionality problem in the feature selection stage [10]. After selecting relevant features, two major challenges have been identified: ensuring data integrity and efficiently handling large datasets using a DL approach [11]. However, the inability to accurately predict patient prognosis prevents pharmacists from providing timely and personalized guidance. DL-based approaches have demonstrated high efficiency in diabetes prediction and management [12]. Consequently, selecting an appropriate classifier becomes crucial for disease diagnosis as it must effectively identify essential features to enhance overall performance [13].

This study utilizes data mining techniques to predict diabetes and employs a DL-based approach to analyze its impact on patients, improving prediction accuracy compared to manual methods. The performance of data classification method is compared using three data containing diabetes and non-diabetes patients [14], [15]. This research conducts studies on diabetes prediction based on DL approach, providing insights into various techniques along with their advantages and limitations. Rastogi and Bansal [16] presented data mining techniques like random forest (RF), support vector machine (SVM), logistic regression (LR), as well as naïve Bayes for diabetes prediction. These techniques support minimizing the risk of overfitting, handling high-dimensional data, and extracting essential features, while SVM maximizes the margin between classes for enhanced performance. However, LR and integrated various techniques require fine-tuning, as processing gastrointestinal (GI) images presents challenges in parameter optimization. Altamimi *et al.* [17] suggested an automated diabetes prediction method of k-nearest neighbor (KNN) imputer and stacked ensemble voting classifier, which effectively addressed the missing data problem. The KNN imputer effectively estimates missing values by leveraging similarities between data points, preserving integrity, and reducing bias. However, it is sensitive to noise, resulting in inaccurate imputation, while stacking multiple models needs careful tuning, and selecting relevant data remains challenging due to the large dataset size. Reza *et al.* [18] introduced an SVM that was highly effective in various classification tasks. The integration of a new kernel with the radial basis function (RBF) with other kernels in SVM enhances its ability to learn complex decision boundaries effectively. The RBF kernel enables the SVM model to learn complex, non-linear decision boundaries through effective intricate patterns and enhances the ability to learn from data with unique distributions. However, SVM struggles with scalability for large data due to quadratic complexity and an inappropriate kernel, leading to suboptimal performance.

Nandeesh and Prasanna [19] presented an ensemble model for diabetes prediction, incorporating RF with grid search, extreme gradient boosting (XGBoost), natural gradient boosting (NGBoost), bagging, light gradient boosting machine (LightGBM), and adaptive boosting (AdaBoost) classifiers. The ensemble models enhance prediction accuracy by aggregating results from multiple models, ensuring robust performance and efficient processing of large datasets. However, the ensemble model often requires substantial computational resources and various model challenges to interpret the decision-making process of individual features. Oliullah *et al.* [20] implemented an Adamic-Adar similarity indexed wald boost data classification (AASIWBDC) approach for the classification and diagnosis of diabetes. An AASIWBDC effectively minimized dimensionality and provided a robust classification framework by leveraging sequential decision-making to optimize accuracy. However, its dependency on feature relationships and the lack of transparency in the boosted classification process make diabetes prediction challenging. Ravichandran *et al.* [21] developed a diabetes prediction approach to assess the risk of readmission using ensemble learning through XGBoost, SVM, and artificial neural network (ANN). By combining multiple models, the approach enhances prediction accuracy while reducing the risk of overfitting in sensitive data. However, the integration of ANN and ensemble methods faces challenges in parameter optimization and struggles to analyze complex patterns effectively. From the overall analysis, the existing techniques face challenges like missing data and struggle to handle high-dimensional data, which impacts diabetes prediction.

The traditional prediction models frequently exclude records with missing values, which leads to valuable patient information loss. This results bias model performance and limits real-world clinical applicability and accuracy. The existing techniques also face high dimensionality issues that impact overall diabetic prediction efficiency.

This research proposes ridgelet process neural network (RPNN) with noise contrastive estimation (NCE) for the classification of diabetes and to mitigate the effects of noise and remove irrelevant features. In

the initial pre-processing stage, standard scaling is involved to ensure data categorization, and missing data is handled using the interquartile range (IQR) method, which imputes values to improve performance during training and prediction. The main contribution of the research is considered as trials:

- i) The ridgelet procedure enhances the network's ability to focus on data structure and handle complex patterns, while its combination with InfoNCE improves model performance and analysis of unaffected diabetes patients.
- ii) The partial least angle regression (PLAR) efficiently minimizes the feature space by selecting relevant features and effectively managing high-dimensional data.
- iii) The RPNN with a ridgelet transformer maps input data to a higher-dimensional space, modelling the large features and handling linear patterns effectively.

This research paper is ordered as follows. Section 2 provides a proposed methodology. Section 3 illustrates the RPNN-NCE for prediction. Section 4 provides the results and discussion. The conclusion is given in section 5.

## 2. PROPOSED METHOD

This research proposes RPNN-NCE for classification of diabetes and to mitigate the effects of noise and remove irrelevant features. In the initial pre-processing stage, standard scaling is involved to ensure data categorization, and missing data is handled using the IQR method, which imputes values to improve performance during training and prediction. The feature selection using PLAR is utilized to efficiently handle high-dimensional data and minimize multicollinearity among features. The RPNN with a ridgelet transformer maps data into higher-dimensional space, enabling the capture of complex features and efficient handling of linear patterns. Additionally, the integration of RPNN-NCE improves a model's capability for identifying relationships through maximizing mutual information. Figure 1 demonstrates the pipeline of the implemented methodology.

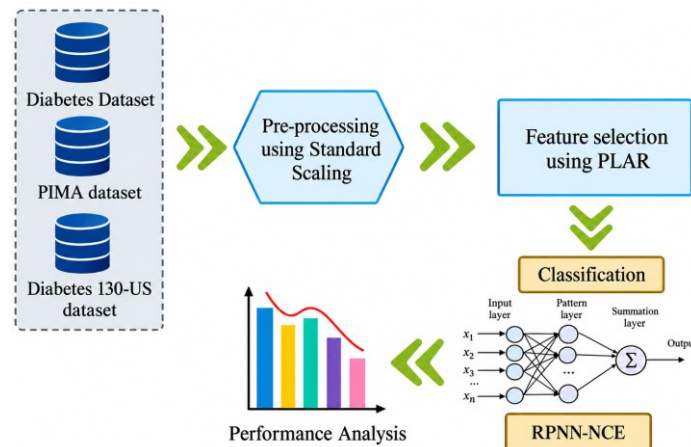


Figure 1. Pipeline of implemented methodology

### 2.1. Dataset

This research considers three datasets: diabetes, PIMA, and diabetes 130-US dataset from the data mining to estimate the effectiveness of the introduced approach. These datasets are widely recognized benchmarks in medical data mining, enabling comprehensive evaluation over large patient records. These datasets offer varying complexities, feature distributions, and class imbalances, providing robust validation for the proposed method.

#### 2.1.1. Diabetes dataset

Diabetes dataset [22] involves 768 instances of female patients aged 21 and older. Among these, 268 cases are diagnosed with diabetes, while the other 500 represent non-diabetic individuals. Notably, the dataset includes 227 individuals with a skinfold thickness of zero, 35 patients with a diastolic blood pressure of zero, and 27 patients with a body mass index (BMI) of zero.

### 2.1.2. PIMA dataset

PIMA Indian diabetes dataset sourced from the UCI ML repository is a widely used dataset for diabetes prediction [23]. It includes data from 786 women, both with and without diabetes, all aged 21 years or older, diagnosed with a type of diabetes. The data comprises nine attributes, 8 of which are related to diabetes diagnosis, including pregnancy, BMI, insulin levels, age, blood pressure, skin thickness, glucose levels, and diabetes pedigree function, along with one label attribute indicating the diabetes status. The dataset includes 268 positive cases and 500 negative cases reflecting the distribution of individuals affected by diabetes.

### 2.1.3. Diabetes 130-US dataset

Diabetes 130-US Hospitals dataset for the years 1999-2008 [24] was sourced from the UCI ML repository. It provides extensive patient information and medical records appropriate for diabetes diagnosis. It includes 55 features and 100,000 instances, converting details like patient identifier, race, age, gender, admission type, length of hospital stay, medical field of admitting physician, and count of lab tests conducted.

## 2.2. Pre-processing

After the data is acquired, the pre-processing is performed by handling missing values and data standardization. These approaches effectively and accurately support early diabetes prediction. The details of these approaches are described in the following.

- i) The missing values significantly affect data accuracy and prediction outcomes, emphasizing the critical role of data preprocessing in ensuring model reliability. Some attributes in the dataset contain incomplete or inaccurate values due to errors or missing data. This has led to deceptive outcomes like diastolic blood pressure, systolic blood pressure, and BMI values being recorded as zero in initial data.
- ii) To handle missing data [25], values are replaced with the median rather than being removed, as missing values negatively impact prediction performance. Median imputation is defined as  $P(x) = \text{median}(x)$ , is an effective method for imputing numerical data. Additionally, the IQR method is used to identify outliers, revealing that the glucose feature contained all the identified outliers. The decision range is defined to help the approach in (1) and (2).

$$\text{Upper bound: } P3 + 1.5 \times IIR \quad (1)$$

$$\text{Lower bound: } P1 - 1.5 \times IIR \quad (2)$$

Where IQR is defined as  $P3 - P1$  and  $P3$  and  $P1$  dataset 3rd represents the third and first quartiles of the dataset, respectively. Dataset outside this range is treated as outliers and eliminated to prevent unequal feature vectors from negatively impacting prediction performance. This is performed by using the Z-score method, which standardizes the data based on its mean and standard deviation.

- iii) Data standardization is performed by using the Z-score method to standardize the data. This approach is based on its mean and standard deviation, ensuring it pursues a normal distribution. The mean 0 and standard deviation 1 are defined by (3).

$$x^* = \frac{x - \mu}{\sigma} \quad (3)$$

Where  $x^*$  denoted the transformed version of the original data  $x$  indicated original data,  $\mu$  denoted mean of data, and  $\sigma$  is standard deviation of data. This preprocessing step enhances model stability, accelerates convergence, and improves the overall accuracy of the prediction algorithm. After pre-processing, the data is fed to the feature selection process to reduce the high dimensionality issue.

## 2.3. Feature selection

After pre-processing, feature selection is performed using PLAR a vector-based algorithm that accelerates computations through a stage-wise approach. PLAR applies the partial least squares (PLS) method to identify parameters that determine the feature attribute. Initially, all coefficients ( $\beta$ ) are set to 0, and the algorithm identifies the covariates which are mostly correlated with the response variable. The process involves partitioning the dataset into complementary subsets, where one subset is used for analysis and the other for validation. The algorithm proceeds by moving in the direction of the covariate with the highest correlation to the residual until another covariate exhibits an equal correlation. This stepwise process continues until the k-th covariate is incorporated into the model. The selected feature denoted  $k$  aims to produce a simpler and more generalized model. The quality of the PLS model is estimated using (4), which considers various parameters, including the coefficient of determination ( $R^2$ ) for each predicted variable.

$$R' = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

The values represent the observed data, while predicted values are derived from the PLS model. The mean values and sample count are also considered and higher values demonstrate as better model. Reflecting the relationship between descriptor variables and dependent variables with an upper limit of 1 for a flawless set and a threshold of 0.5 for an acceptable fit. This metric quantifies the extent to which the model replicates observed outcomes and determines the proportion of variance in a reliant on variable which is predicted from self-determining variables. The cross-validation  $R^2$  represents the predictive accuracy for internal validation when the corresponding determination coefficient for external validation measures performance on unseen data. The root mean square error of calibration (RMSEC) quantifies a variation among experimental as well as predicted values in the PLS model. A lower RMSEC indicates better model predictions. Additionally, RMSEC values for cross-validation and external validation are expressed by (5) and (6). The average overall error was defined by (6) to relate RMSEC respective equation.

$$RMSEC = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)}{n} \quad (5)$$

$$Average\ OE = \frac{100 \times RMSE}{y\theta_i} \quad (6)$$

Where  $OE_{CV}$  and  $OE_{ext}$  are evaluated values for cross-validation and external validation gave information on relative precision. The diabetes disease quality of the model informs on the unexplained variance, enabling the identification of diabetes diseases in the model. This approach efficiently selects the most relevant features by minimizing multicollinearity and reducing dimensionality.

### 3. PREDICTION USING RPNN-NCE

The RPNN is a specialized neural network that utilizes ridgelet transformations to map input data into higher-dimensional space, enhancing a network's ability to capture non-linear relationships. This approach is particularly effective for sparse datasets like diabetes data that continuously display complex and non-linear relationships among features. An RPNN approach addresses mapping approximation challenges caused by sudden continuous variations in asymmetric operating curves by utilizing the ridgelet transformation as the activation function of the neural network. Figure 2 illustrates the structure of the RPNN technique.

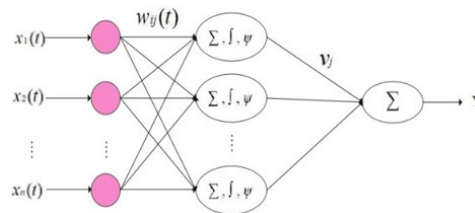


Figure 2. Structure of the RPNN technique

The connection weight function  $W(t) = \{w_1(t), w_2(t), \dots, w_n(t)\}$ ,  $\psi$  denoted the ridgelet function and a hidden layer RPNN structure. A relationship between the input and output of RPNN is expressed in (7). A group of orthogonal basis functions  $b_1(t), b_2(t), \dots, b_L(t)$  is chosen as the input function space, which is expressed in (8). The accuracy of input and weight function fitting is determined based on (7) and (8), while (9) is the property of an orthonormal basis in (10).

$$y = \sum_{j=1}^m v_j \psi \left( \left( \frac{u_j \int_0^T \sum_{i=1}^n w_{ij}(t) x_i(t) dt - b_j}{a_j} \right) \right) \quad (7)$$

$$x_i(t) = \sum_{l=1}^L b_l^{(i)} b_{ij}(t) = \sum_{k=1}^L w_{ik} b_k(t) \quad (8)$$

Where  $w_{ij}(t)x_i(t)$  denoted inner product of  $i$ th input  $x_i(t)$  and  $w_{ij}(t)$ .

$$y = \sum_{j=1}^m v_{jj} \Psi \left| \frac{\left| \frac{u_j \sum_{l=1}^n \nabla_{l=1}^L \nabla_{k=1}^L \int_0^b (t)^{(k)}(t) dt - b}{\int_0^{(k)} \frac{(k)}{l} \frac{(l)}{l}} \right|}{\frac{a_{jj}}{h}} \right| \quad (9)$$

$$E = \frac{1}{2} \sum_{s=1}^K (y_k - d_k)^2 = \frac{1}{2} \sum_{s=1}^K \left| \sum_{j=1}^m \left( \frac{u_{ij} \sum_{l=1}^n \nabla_{l=1}^L v_{ij} \psi(u_i(u_i(k) - h..)}{a_{ji}} \right) - d_k \right| \quad (10)$$

If the sample is  $(x_{k1}(t), x_{k2}(t), \dots, x_{kn}(t), d_k)$ ,  $k = 1, 2, \dots, K$  an error function of RPNN is expressed in (10). The RPNN technique faces challenges in analyzing nonlinear patterns and fails to capture relationships between layers. To address this, InfoNCE loss is used to estimate layer relationships. Contrastive learning helps by reducing the distance among positive pairs whereas increasing the separation among negative pairs into depiction space. The purpose of this method forces an approach for significantly classifying positive pairs from their negative complements and lack of loss of generalization deliberate NCE-based contrastive learning approach. The anchor  $x$ , positive sample  $x^+$  and  $m$  negative samples  $\{\underline{x}_{ij}^-\}^m$  model needed to reduce an InfoNCE loss defined in (11).

$$\mathcal{L} = \mathbb{E}_{x, x^+, \{x^-\}_{i=1}^m} \left[ -\log \frac{e^{\frac{s(z, z^+)}{\tau}}}{e^{\frac{s(z, z^+)}{\tau}} + \sum_{i=1}^m e^{\frac{s(z, z^+)}{\tau}}} \right] \quad (11)$$

The RPNN with the NCE function distinguishes healthy patients from diabetic patients while efficiently learning complex patterns from the dataset. It analyzes both linear and nonlinear patterns to identify diabetes cases, utilizing a process neural network that integrates with the ridgelet neural network to handle specific layers effectively. The RPNN outperforms traditional neural networks by ensuring stable data conversions, enabling faster parallel processing and improved performance. The RPNN method using ridgelet transforms to extract local geometric features which handles high dimensionality, significantly decreases complex inputs into manageable patterns. The NCE identifies missing data thereby model training to differentiate correct patient records from the synthetic noise, enhancing better prediction even with incomplete information. The integration of both RPNN and NCE method improves robust prediction of diabetes by mapping complex variables and reconstructing incomplete clinical records into accurate, lower-dimensional representations [26], [27].

#### 4. RESULTS AND DISCUSSION

The proposed method is simulated using Python with Windows 10 OS, 8 GB RAM, and an i5 processor. The PIMA, diabetes, and diabetes 130-US datasets are considered to estimate the performance of the proposed RPNN-NCE method. The all three dataset dataset splitting is provided based on the training, testing, and validation sets, and class details. The model learns underlying patterns of features by isolating training, and testing provides reliable model tuning. Table 1 demonstrates the splitting of PIMA, diabetes, and diabetes-130 US datasets under which scenario the simulated results are evaluated.

Table 1. The dataset splitting with respect to the dataset

| Dataset           | Total records | Training samples  | Testing samples (%) | Validation (%) | Classes                                          |
|-------------------|---------------|-------------------|---------------------|----------------|--------------------------------------------------|
| PIMA              | 768           | PIMA              | 20                  | 10             | Binary [0: Negative, 1:Positive]                 |
| Diabetes – 130 US | 101,766       | Diabetes – 130 US | 15                  | 15             | Muulticlass [No readmission, <30 days, >30 days] |
| Diabetes          | 100,000       | Diabetes          | 15                  | 15             | Binary [0: No Diabetes, 1: Diabetes]             |

The metrics like accuracy, precision, recall, F1-score, and area under the curve (AUC) are estimated to assess the performance. Their mathematical formulas are given in (12) to (16). Where  $TP$  is true positive,  $TN$  is true negative,  $FP$  is false positive, and  $FN$  is false negative.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (12)$$

$$Precision = \frac{TP}{TP+FP} \times 100 \quad (13)$$

$$Recall = \frac{TP}{TP+FN} \times 100 \quad (14)$$

$$F1 - score = \frac{2 \times Precision \times recall}{Precision + recall} \quad (15)$$

$$MissRate = \frac{(FP+FN)}{(TP+TN+FP+FN)} \quad (16)$$

#### 4.1. Performance analysis

Figure 3 demonstrates the performance of the dimensional reduction process on the diabetes dataset. Figure 4 displays a performance analysis of dimensional reduction on the PIMA dataset. Figure 5 illustrates the performance analysis of dimensional reduction on the diabetes 130-US dataset. The performance analysis of RPNN-NCE is estimated and compared with various existing methods based on standard datasets. The PLAR is compared and evaluated with other feature selection methods like recursive feature elimination (RFE), Chi-square, Lasso L1, and principal component analysis (PCA) to show the model's effectiveness. The PLAR method reaches greater accuracy of 98.94%, precision of 97.62%, recall of 96.02%, and F1-score of 96.79%, respectively. The proposed method is compared and evaluated with various existing methods like deep neural network (DNN), CNN, ANN, and gated recurrent unit (GRU) based on the standard datasets to show model performance.

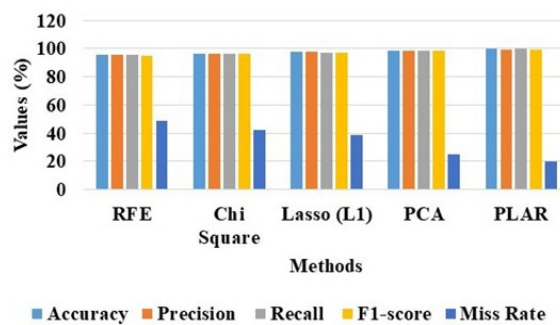


Figure 3. Performance of the dimensional reduction process on the diabetes dataset

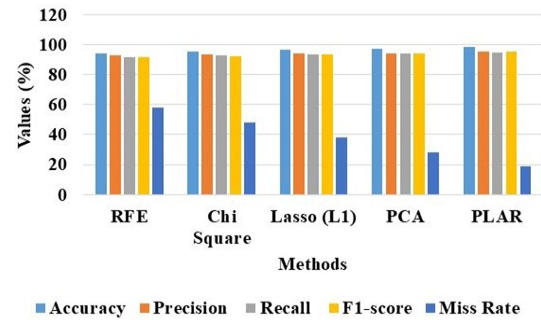


Figure 4. Performance analysis of the dimensional reduction on PIMA dataset

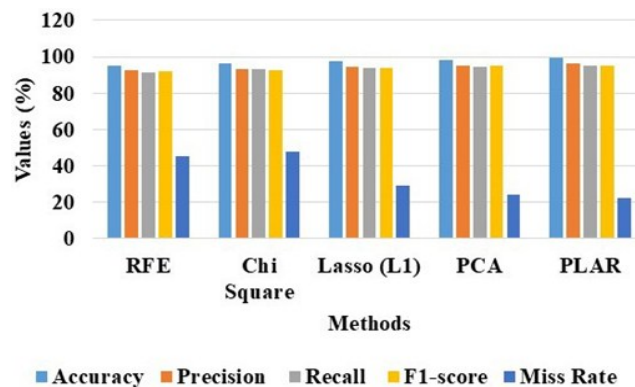


Figure 5. Performance analysis of dimensional reduction on diabetes 130-US datasets

Table 2 represents the performance analysis by comparing the proposed method with existing approaches with respect to the PIMA, diabetes, and diabetes 130-US datasets. The proposed RPNN-NCE method is compared with the traditional DNN, GRU, ANN, and CNN by considering the performance metrics. The RPNN-NCE method achieves higher accuracy when compared with the other DL techniques to improve the overall model efficiency.

Table 2. Performance analysis of classification methods

| Datasets                 | Methods  | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) | Miss rate (%) |
|--------------------------|----------|--------------|---------------|------------|--------------|---------------|
| Diabetes dataset         | ANN      | 95.63        | 95.78         | 95.75      | 95.12        | 49            |
|                          | CNN      | 96.25        | 96.45         | 96.35      | 96.35        | 42            |
|                          | GRU      | 97.85        | 97.63         | 97.12      | 97.03        | 39            |
|                          | DNN      | 98.63        | 98.26         | 98.41      | 98.23        | 25            |
|                          | RPNN-NCE | 99.89        | 99.28         | 99.85      | 99.56        | 20            |
| PIMA dataset             | ANN      | 94.08        | 92.63         | 91.85      | 91.64        | 58            |
|                          | CNN      | 95.47        | 93.62         | 92.75      | 92.35        | 48            |
|                          | GRU      | 96.35        | 94.12         | 93.65      | 93.62        | 38            |
|                          | DNN      | 97.03        | 94.36         | 94.35      | 94.05        | 28            |
|                          | RPNN-NCE | 98.63        | 95.23         | 95.03      | 95.12        | 19            |
| Diabetes 130-US datasets | ANN      | 95.23        | 92.85         | 91.03      | 91.75        | 45            |
|                          | CNN      | 96.43        | 93.03         | 92.87      | 92.36        | 48            |
|                          | GRU      | 97.68        | 94.68         | 93.65      | 93.65        | 29            |
|                          | DNN      | 98.36        | 95.23         | 94.26      | 94.85        | 24            |
|                          | RPNN-NCE | 99.37        | 96.53         | 95.23      | 95.23        | 22            |

#### 4.2. Statistical analysis

The proposed method provides a statistically significant reduction in classification errors, which is compared with other existing methods such as DNN, CNN, GRU, and ANN respectively. The proposed method p-value shows the generalization ability based on the analyzing statistically. Table 3 demonstrates the McNemar statistical analysis based on the parameter accuracy repeated across 5 runs in similar data split conditions.

Table 3. The statistical analysis by considering the McNemar score

| Model             | Accuracy (%) | Expected discordant pairs (b>c) | McNemar $\chi^2$ Score | p-value |
|-------------------|--------------|---------------------------------|------------------------|---------|
| ANN               | 82.1         | High                            | 50.0                   | 0.048   |
| CNN               | 96.2         | Moderate                        | 20.0                   | 0.037   |
| GRU               | 94.0         | Moderate                        | 25.0                   | 0.032   |
| DNN               | 95.5         | Moderate                        | 15.0                   | 0.025   |
| Proposed RPNN-NCE | 99.89        | NA                              | NA                     | 0.018   |

#### 4.3. Comparative analysis

In this section, the proposed method's effectiveness is compared with existing methods like LR [16], ensemble voting classifier (XGBoost + RF + extra tree classifier or ETC) [17], SVM kernel [18], ensemble model [19], AASIWBDC [20], and guided artificial neural network (G-ANN) [21] using various standard datasets. The proposed method achieves better accuracy of 99.89%, 98.63%, and 99.37% on diabetes, PIMA, and diabetes 130 US datasets. Table 4 demonstrates the comparative analysis of the proposed method.

Table 4. Comparative analysis of the proposed method on datasets

| Datasets                 | Methods                    | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|--------------------------|----------------------------|--------------|---------------|------------|--------------|
| Diabetes dataset         | [16]                       | 82.46        | NA            | 68.23      | NA           |
|                          | [17]                       | 98.59        | 99.26         | 99.75      | 99.45        |
|                          | Proposed RPNN – NCE method | 99.89        | 99.28         | 99.85      | 99.56        |
| PIMA dataset             | [18]                       | 85.5         | NA            | NA         | 85.5         |
|                          | [19]                       | 92.91        | 93            | 93         | 92           |
|                          | Proposed RPNN – NCE method | 98.63        | 95.23         | 95.03      | 95.12        |
| Diabetes 130-US datasets | [20]                       | 98.99        | NA            | NA         | NA           |
|                          | [21]                       | 89.16        | NA            | NA         | NA           |
|                          | Proposed RPNN – NCE method | 99.37        | 96.53         | 95.23      | 95.23        |

#### 4.4. Discussion

The advantages of proposed method and limitations of existing methods like LR [15] and combining multiple algorithms require extensive fine-tuning for processing GI images. The KNN imputation [17] is highly sensitive to noise, leading to inaccurate imputations, while ensemble models demand careful tuning and pose challenges in data selection due to the complexity of large datasets. The SVM [18] struggle with scalability for large data due to quadratic complexity and an inappropriate kernel, leading to suboptimal performance. The ensemble model [19] often requires substantial computational resources and various models challenging to interpret the decision-making process of individual features. The AASIWBDC [20] depends on the essential relationships between features, however, the boosted classification process lacks transparency, making it challenging to predict diabetes accurately. The G-ANN [21] and ensemble method integrating the prediction of diabetes patients struggle to process the parameters and are unable to analyze the complex patterns. Additionally, the integration of NCE with RPNN enhances the model's ability to identify relationships by maximizing mutual information.

#### 5. CONCLUSION

This research proposes the RPNN-NCE for the classification of diabetes and to mitigate the effects of noise and remove irrelevant features. In the initial pre-processing stage, standard scaling is involved to ensure data categorization, and missing data is handled using the IQR method, which imputes values to improve performance during training and prediction. The feature selection using PLAR is utilized to efficiently handle high-dimensional data and minimize multicollinearity among features. The RPNN, utilizing a ridgelet transformer, maps data into a higher-dimensional space, enabling the capture of complex features and efficient handling of linear patterns. Additionally, the integration of RPNN-NCE improves the network's capability to identify relationships by maximizing mutual information. The proposed method achieves better accuracy of 99.89%, 98.63%, and 99.37% on datasets like diabetes, PIMA, and diabetes 130-US when compared with existing methods like G-ANN. Future work will consider the hybrid DL approach to effectively predict diabetes by using various datasets.

#### ACKNOWLEDGMENTS

The authors thank the faculty and research scholars of the Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, JSS Science and Technology University Mysuru, and the Department of Data Science, Dayanand Sagar Academy of Technology and Management, Bengaluru, for their valuable guidance and support.

#### FUNDING INFORMATION

Authors state no funding involved.

#### AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author        | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|-----------------------|---|---|----|----|----|---|---|---|---|---|----|----|---|----|
| Parashiva Murthy      | ✓ | ✓ | ✓  | ✓  | ✓  | ✓ |   | ✓ | ✓ | ✓ |    |    | ✓ |    |
| Basavanapura          |   |   |    |    |    |   |   |   |   |   |    |    |   |    |
| Muddumadappa          |   |   |    |    |    |   |   |   |   |   |    |    |   |    |
| Pruthvi Palahalli     |   | ✓ |    |    |    | ✓ |   | ✓ | ✓ | ✓ | ✓  | ✓  |   |    |
| Ramachandra           |   |   |    |    |    |   |   |   |   |   |    |    |   |    |
| Manjunatha Basavaraju | ✓ |   | ✓  | ✓  |    |   | ✓ |   |   | ✓ | ✓  |    | ✓ | ✓  |
| Nandeesh Hallimysore  |   |   |    |    | ✓  |   | ✓ |   |   | ✓ |    | ✓  |   | ✓  |
| Devaraj               |   |   |    |    |    |   |   |   |   |   |    |    |   |    |

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nterpretation

R : **R**esources

D : **D**ata Curation

O : **O**riginal Draft

E : **E**diting

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## DATA AVAILABILITY

The data that support the findings of this study are openly available diabetes dataset in <https://archive.ics.uci.edu/dataset/34/diabetes> [22], [26], PIMA dataset in <https://www.openml.org/search?type=data&id=43483&sort=runs&status=active> [23], and diabetes 130-US dataset in <https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008> [24], [27].

## REFERENCES

- [1] J. K. Shimpi, P. Shanmugam, and A. A. Stonier, "Analytical model to predict diabetic patients using an optimized hybrid classifier," *Soft Computing*, vol. 28, no. 3, pp. 1883–1892, Dec. 2024, doi: 10.1007/s00500-023-09487-w.
- [2] S. K. S. Modak and V. K. Jha, "Diabetes prediction model using machine learning techniques," *Multimedia Tools and Applications*, vol. 83, no. 13, pp. 38523–38549, Oct. 2024, doi: 10.1007/s11042-023-16745-4.
- [3] S. S. Gupta *et al.*, "Diabetes estimation through data mining using optimization, clustering, and secure cloud storage strategies," *SN Computer Science*, vol. 5, no. 6, Aug. 2024, doi: 10.1007/s42979-024-03158-9.
- [4] J. Abdollahi and S. Aref, "Early prediction of diabetes using feature selection and machine learning algorithms," *SN Computer Science*, vol. 5, no. 2, Jan. 2024, doi: 10.1007/s42979-023-02545-y.
- [5] J. Xiao *et al.*, "Interpretable time-series neural turing machine for prognostic prediction of patients with type 2 diabetes in physician-pharmacist collaborative clinics," *International Journal of Medical Informatics*, vol. 195, Mar. 2025, doi: 10.1016/j.ijmedinf.2024.105737.
- [6] P. Sahu and J. K. Mantri, "Performance evaluation of data mining and neural network based models for diabetes prediction," in *2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing, ICSTSN 2023*, Apr. 2023, pp. 1–10, doi: 10.1109/ICSTSN57873.2023.10151474.
- [7] M. Zeinalnezhad and S. Shishehchi, "An integrated data mining algorithms and meta-heuristic technique to predict the readmission risk of diabetic patients," *Healthcare Analytics*, vol. 5, Jun. 2024, doi: 10.1016/j.health.2023.100292.
- [8] O. Y. Dwekat and S. S. Lam, "Optimized design of hybrid genetic algorithm with multilayer perceptron to predict patients with diabetes," *Soft Computing*, vol. 27, no. 10, pp. 6205–6222, Feb. 2023, doi: 10.1007/s00500-023-07876-9.
- [9] G. Dharmarathne, T. N. Jayasinghe, M. Bogahawaththa, D. P. P. Meddage, and U. Rathnayake, "A novel machine learning approach for diagnosing diabetes with a self-explainable interface," *Healthcare Analytics*, vol. 5, Jun. 2024, doi: 10.1016/j.health.2024.100301.
- [10] M. A. Uddin *et al.*, "Machine learning based diabetes detection model for false negative reduction," *Biomedical Materials and Devices*, vol. 2, no. 1, pp. 427–443, Jun. 2024, doi: 10.1007/s44174-023-00104-w.
- [11] P. Whig, K. Gupta, N. Jiwani, H. Jupalle, S. Kouser, and N. Alam, "A novel method for diabetes classification and prediction with Pycaret," *Microsystem Technologies*, vol. 29, no. 10, pp. 1479–1487, Jun. 2023, doi: 10.1007/s00542-023-05473-2.
- [12] M. S. Al Reshan *et al.*, "An innovative ensemble deep learning clinical decision support system for diabetes prediction," *IEEE Access*, vol. 12, pp. 106193–106210, Aug. 2024, doi: 10.1109/ACCESS.2024.3436641.
- [13] H. Zhou, Y. Xin, and S. Li, "A diabetes prediction model based on Boruta feature selection and ensemble learning," *BMC Bioinformatics*, vol. 24, no. 1, Jun. 2023, doi: 10.1186/s12859-023-05300-5.
- [14] M. Khalifa and M. Albadawy, "Artificial intelligence for diabetes: enhancing prevention, diagnosis, and effective management," *Computer Methods and Programs in Biomedicine Update*, vol. 5, Jan. 2024, doi: 10.1016/j.cmpbup.2024.100141.
- [15] F. A. Jaber and J. W. James, "Early prediction of diabetic using data mining," *SN Computer Science*, vol. 4, no. 2, Jan. 2023, doi: 10.1007/s42979-022-01594-z.
- [16] R. Rastogi and M. Bansal, "Diabetes prediction model using data mining techniques," *Measurement: Sensors*, vol. 25, Feb. 2023, doi: 10.1016/j.measen.2022.100605.
- [17] A. Altamimi *et al.*, "An automated approach to predict diabetic patients using KNN imputation and effective data mining techniques," *BMC Medical Research Methodology*, vol. 24, no. 1, 2024, doi: 10.1186/s12874-024-02324-0.
- [18] M. S. Reza, U. Hafsha, R. Amin, R. Yasmin, and S. Ruhi, "Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: insights from the PIMA dataset," *Computer Methods and Programs in Biomedicine Update*, vol. 4, Jan. 2023, doi: 10.1016/j.cmpbup.2023.100118.
- [19] H. D. Nandeeshha and B. T. Prasanna, "Enhanced phishing URL classification based on machine learning with keyword specific web crawler," in *Advances in Communication and Applications (ERCICA 2024)*, Singapore: Springer, 2025, pp. 443–456, doi: 10.1007/978-981-96-4679-1\_35.
- [20] K. Oliullah, M. H. Rasel, M. M. Islam, M. R. Islam, M. A. H. Wadud, and M. Whaiduzzaman, "A stacked ensemble machine learning approach for the prediction of diabetes," *Journal of Diabetes and Metabolic Disorders*, vol. 23, no. 1, pp. 603–617, Nov. 2024, doi: 10.1007/s40200-023-01321-2.
- [21] M. Ravichandran *et al.*, "Adamic–Adar similarity indexed wald boost data classification for diabetic disease diagnosis with big data," *Systems and Soft Computing*, vol. 6, Dec. 2024, doi: 10.1016/j.sasc.2024.200175.
- [22] A. A. Ram, Z. Ali, V. Krishna, N. Nishika, and A. Sharma, "A guided neural network approach to predict early readmission of diabetic patients," *IEEE Access*, vol. 11, pp. 47527–47538, May 2023, doi: 10.1109/ACCESS.2023.3275086.
- [23] M. Kahn, "Diabetes [Dataset]," *UCI Machine Learning Repository*, 2025, doi: 10.24432/C5T59G.
- [24] O. Yildirim, "PIMA dataset," *openml.org*, 2022. Accessed: Dec. 21, 2025. [Online]. Available: <https://www.openml.org/search?type=data&id=43483&sort=runs&status=active>
- [25] J. Clore, K. Cios, J. DeShazo, and B. Strack, "Diabetes 130-US hospitals for years 1999-2008 [Dataset]," *UCI Machine Learning Repository*, 2025, doi: 10.24432/C5230J.
- [26] A. Palanivayagam and R. Damaševičius, "Effective handling of missing values in datasets for classification using machine learning methods," *Information*, vol. 14, no. 2, Feb. 2023, doi: 10.3390/info14020092.
- [27] N. H. Devaraj, P. B. Thimappa, and P. Shivarajappa, "Phish-view: a feature-rich meta ensemble learning for accurate and uncertainty-aware phishing detection," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 7, pp. 190–209, 2025, doi: 10.22266/ijies2025.0831.14.

## BIOGRAPHIES OF AUTHORS



**Parashiva Murthy Basavanapura Muddumadappa**    is a dedicated professional in the field of Computer Science and Engineering, who is currently working as an assistant professor in the Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, JSS Science and Technology University, Mysuru, making remarkable contributions to Academia and Research. His journey in the field of academic landscape began at JSS Science and Technology University, where he initiated her career excellence that would span various roles and achievements. He has a cumulative experience of 12 years and has published several research papers in reputed journals and conferences. Education has always been at the forefront. He has been instrumental in delivering several faculty development programs associated with the technical education quality improvement program of the Government of India (TEQIP). He works in the areas of web development, business intelligence, and data analytics. He can be contacted at email: parashivamurthy@sjce.ac.in.



**Pruthvi Palahalli Ramachandra**    is an assistant professor in the Department of Computer Science and Engineering at Vidyavardhaka College of Engineering, Mysuru, Karnataka, India. She obtained her B.E. and M.Tech. degrees in Computer Science and Engineering from Visvesvaraya Technological University, Belagavi. She is currently pursuing her Ph.D. at the research center, Vidyavardhaka College of Engineering, Mysuru, affiliated with Visvesvaraya Technological University, Belagavi, with a research focus on medical imaging. Her research interests include medical imaging, artificial intelligence, machine learning, deep learning, internet of things, and data science. She has authored and co-authored several research papers published in IEEE conferences, Springer conference proceedings, and Taylor & Francis journals indexed in Scopus. Her research focuses on developing intelligent, data-driven solutions for real-world engineering and healthcare applications. She can be contacted at email: pruthvi.pr@vvce.ac.in.



**Manjunatha Basavaraju**    is an associate professor at the Department of Industrial and Production Engineering, Sri Jayachamarajendra College of Engineering, JSS Science and Technology University, Mysuru, Karnataka, India. He has extensive experience in teaching, research, and academic administration. His academic interests encompass advanced manufacturing, industrial production engineering, smart manufacturing, automation, Industry 4.0/5.0 technologies, production planning and control, quality engineering, sustainable manufacturing systems, and AI applications. He can be contacted at email: manjub@jssstuniv.in.



**Nandeesh Hallimysore Devaraj**    is an assistant professor in the Department of Computer Science and Engineering, JSS Science and Technology, Mysuru, Karnataka, India. He received a Master of Technology from Sri Jayachamarajendra College of Engineering. Currently, he is pursuing Ph.D. in cyber security JSS Science and Technology University, Mysuru. His general research interest is in the area of information and cyber security, URL phishing detection, web security, mobile security, online social networks, and machine learning. He can be contacted at email: hdnandeesh@sjce.ac.in.