# PREDICTING THE ONSET OF DIABETES USING RNN-RBFN: A ROBUST DEEP LEARNING MODEL FOR PRECISION IN HEALTHCARE

**S. Gomathi[1], B. Dhiyanesh[2*], J.K. Kiruthika[3], P. Saraswathi[4], K. Divya[5], R. Sangeethapriya[6],**

[1]Assistant Professor, Department of Computer Science Engineering, Dr. N.G.P Institute of Technology, Coimbatore,Tamil Nadu 641048, India.
**Email: mail2mathi86@gmail.com**

[2*]Associate Professor, Department of Computer Science Engineering, SRM Institute of Science and Technology, Chennai,Tamil Nadu 600026, India.
**Email:dhiyanu87@gmail.com***

[3]Assistant Professor (Sr. G), Department of Computer Science Engineering, KPR Institute of Engineering and Technology, Coimbatore,TamilNadu 641407, India.
**Email: kiruthika.jk@kpriet.ac.in**

[4]Assistant Professor, Department of Information Technology, Velammal College of Engineering and Technology, Madurai,Tamil Nadu 625009, India.
**Email:psaraswathimtech@gmail.com**

[5]Assistant Professor, Department of Electrical and Electronics Engineering, Karpagam Institute of Technology, Coimbatore,Tamil Nadu 641021, India.
**Email: divya.eee@karpagamtech.ac.in**

[6]Assistant professor, Department of Information Technology, Sona college of Technology, Salem, Tamil Nadu 636005, India.
**Email: priyabecse49@gmail.com**

## Abstract

According to the International Diabetes Federation (IDF), diabetes is a chronic condition that is just one of the glaringly increasing global health issues of our day. Unprocessed diabetes can harm the kidneys, nerves, and heart and initiate eye illnesses like diabetic retinopathy. The healthcare cost of diabetes was 760.3 billion in 2019 and is projected to upsurge to 824.7 billion by 2030, according to IDF. Identifying people at risk can counteract health issues, improve the

superiority of life, and hold back many costs. In healthcare systems, the wrong assortment of lesser groups of patients is costlier than the misclassification of healthy individuals. However, ML algorithms assume a uniform misclassification error and a symmetric class distribution. It is often difficult to analyze diabetes data to predict the onset of diabetes. A Recurrent Neural Network with Radial Basis Function Networks (RNN-RBFN) method based on DL methodology is proposed to resolve the issue. Initially, we collected the diabetic dataset at the Mendeley Data website, which includes a healthcare margin. The proposed method has three stages: preprocessing, feature selection, and classification. In the first stage, the diabetic dataset will be preprocessed based on the Imputation method. It is used to swap the absent data in the dataset with alternative characteristics so that the maximum of the data or info in the dataset is retained. This method is used because it is impractical to remove data from the dataset every time and can significantly reduce its size. The next stage is featuring selection, which uses linear discriminant analysis (LDA) to plot the data in a low-dimensional space to maximize class separation. This is accomplished by finding a set of linear discriminants that maximizes the proportion of between-class variance to within-class variance. The last stage is classification; RNN-RBFN can learn more complex patterns and improve classification performance compared to using either network in isolation. Investigational outcomes determine that our method beats the previous method in precision, recall, F1 score, accuracy, and time complexity.

**Keywords:** Diabetes Prediction, Neural Networks, Radial Basis Function Networks, Healthcare Data, Imputation Method, Linear Discriminant, Diabetic Retinopathy

## I. INTRODUCTION

Globally, diabetes is the foremost basis of escapable blindness in grown people between 20 and 74. Major companies recommend regular testing for early detection. Routine screening is essential to prevent blindness, but as the amount of public with diabetes is predicted to enlarge on or after 415 million in 2015 toward 642 million in 2040, the incumbrance of screening and continuation examinations is becoming a more significant burden[1]. In order to effectively manage diabetic patients, a profound consideration of its antecedent features, premature analysis, and appropriate therapeutic intercession is essential. Initial detection of patients at the hazard of emergent diabetes is significant to active intercession, which is critical to slowing the diabetes movement and, in so doing, tumbling the hazard of blindness. In addition, discrete patients can receive more appropriate treatment based on their risk status. However, there are no out-of-the-ordinary signs in the premature stages of the illness.

[1] Ashish Bora, MS, et al., "Predicting the risk of developing diabetic retinopathy using deep learning", Vol 3 Issue1.

Most people with diabetes do not pursue medical assessment until the disease progresses to the advanced phase, resulting in irrevocable vision loss. Consequently, methods are urgently needed to predict diabetes risk [2] accurately.

Machine learning (ML) methodologies can model complex nonlinear patterns to identify at-risk individuals. Medical researchers are particularly interested in predicting rare diseases compared to the general population. Thus, class unevenness is a communal problem in most clinical datasets. When there is class inequality, the number of cases of minority classes is significantly less than other classes. In healthcare applications, misclassification of minority groups of patients is costlier than misclassification of healthy individuals. However, static learning algorithms always assume a uniform misclassification error and asymmetric class distribution. It is often difficult to analyze diabetes data to predict the onset of diabetes [3].

A Recurrent Neural Network with Radial Basis Function Networks (RNN-RBFN) method based on Deep Learning (DL) methodology was proposed to resolve the issue. RNNs are suitable for processing continuous data and modeling the temporal dependence of diabetes risk factors over time.

On the other hand, RBFN is known for its ability to approximate complex nonlinear functions, which helps capture complex relationships among various risk factors. By combining these two neural network architectures, we aim to enhance each of their strengths to improve the exactness and robustness of diabetes hazard forecast methods. Train and evaluate our proposed methods using a diabetes dataset with features related to diabetes risk factors. The process is trained to foresee a person's likelihood of developing diabetes based on these factors, providing valuable insights for early intervention and personalized clinical strategies. Overall, this novel proposes an RNN-RBFN integration for diabetes risk prediction and subsidizes the rising body of research on the use of DL in healthcare applications.

## A. The main contributions of the RNN-RBFN method are mention below

- This method combines RNN and RBFN, the temporal information of continuous data, and the nonlinear relationship between features, allowing a more comprehensive understanding of diabetes risk factors.

---

[2] Zhao Y, Li X, Li S, et al., Using Machine Learning Techniques to Develop Risk Prediction Models for the Risk of Incident Diabetic Retinopathy Among Patients with Type 2 Diabetes Mellitus,2022.

[3] Sadeghi, S., Khalili, D., Ramezankhani, A. et al. Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods. BMC Med Inform Decis Mak 22, 36 (2022). https://doi.org/10.1186/s12911-022-01775-z

- The method used improves the accuracy of diabetic risk prediction. This is because the model utilizes both methods to capture the dataset's complexity better.
- The RNN-RBFN method, with its potential to significantly advance diabetes risk prediction, provides a powerful framework. By leveraging the complementary advantages of RNN and RFBN, it yields more accurate definitive results, inspiring further research in the field.

## II. LITERATURE SURVEY

Laila, U. E et al. (2021) discussed that the hospitals currently have basic information systems; the large amounts of data generated by these systems must be converted into relevant and valuable info and cannot be cast off to sustain clinical executives. An ensemble learning methodology was deployed to resolve the issue. The deployed methodology collated multiple techniques to estimate bias and variance and improve predictions into an efficient prediction system.

Li, Jet al. (2021) discussed that people with diabetes who do not receive early diagnosis and standard treatment are at risk for severe, multifaceted system tricky situations that can be severe. Since prediabetes is an inevitable phenomenon in the manifestation and expansion of the disease, timely detection and intervention of prediabetes is crucial for the prevention of diabetes. A TCM tongue diagnosis methodology collated with ML protocol was deployed to attain the objective. The deployed methodology's main drawback was that the accuracy prediction needed to be higher and feasible.

Ahmed, Net al. (2021) discussed the objective of finding an effective ML-based classification model to diagnose diabetes in individuals using clinical data. To attain the target, an ML-based classification methodology was deployed. The main drawback of the deployed methodology was that it could not accurately and efficiently identify people with diabetes.

Doğru, Aet al. (2023) discussed that recent advances in integrated ML techniques are essential in the premature diagnosis of diabetes. The main goal is to diagnose diabetes early. To attain this goal, a hybrid super ensemble learning methodology was deployed. The deployed methodology cannot effectively detect diabetic patients and has a low statistical score, which is not convincing.

_____-

[4] Laila, U. E., Mahboob, K, et al., An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning.
[5] Li, J., Chen,et al., International Journal of Medical Informatics, Vol 149.
[6] Ahmed, N., et al., International Journal of Cognitive Computing in Engineering, Vol 2.
[7] Doğru, A., Buyrukoğlu, S. et al., A hybrid super ensemble learning model for the early-stage prediction of diabetes risk, Vol 61.

Tan, Yet al. (2022) discussed that early diabetes risk prediction helps doctors and patients focus on the disease as early as possible, intervene, and effectively reduce the risk of complications. A GA-EL methodology was deployed to attain the objective. The deployed methodology overlay has high prediction accuracy. In addition, GA stacking's strong generalizability and high predictive ability were also validated in early diabetes risk prediction.

Tawfik Beghriche et al. (2021) discussed that as the number of deaths due to diabetes increases yearly, the need to develop an effective system to diagnose diabetes has become inevitable. A well-organized medical decision mechanism was deployed based on the Deep Neural Network (DNN) methodology to resolve the issue. Proposition skills combined with clinical knowledge can increase efficiency, adaptability, and transparency in decision-making. This will reduce the time, energy, and effort of clinical services and improve the accuracy of the final results.

Jian, Yet al. (2021) discussed the main drawback of dealing with missing values and unbalanced data for patients with diabetes. A ML protocol was deployed to resolve the issue. The deployed protocol used different sets of attributes to achieve this. However, the suggested method can also build a good classifier using selected features.

Alex, S.Aet al. (2022) discussed that diabetic examination has fascinated the investigation communal to address approximately absent value and class inequality matters. The effectiveness of classifying diabetes using machine learning techniques is relatively low. A Deep 1D-CNN methodology was deployed to resolve the issue. The deployed method reduces the stimulus of inequity class on forecast execution via a discerning set of assessment gauges.

Fayaz et al. (2022) discussed that in previous works, a variety of machine learning-based detection algorithms have been developed to predict diabetes from the provided dataset. Nevertheless, it has certain drawbacks, including hard to grasp aspects, lengthy testing and training periods, overfitting, and incorrect outputs. Thus, the goal of the proposed research project is to put a number of data mining approaches into practice in order to create an automated and effective diabetes diagnosis system.

---

[8] Tan, Y., Chen, H., et al., Early Risk Prediction of Diabetes Based on GA-Stacking. Applied Sciences, vol 12.

[9] Tawfik Beghriche, et al., "An Efficient Prediction System for Diabetes Disease Based on Deep Neural Network", 2021.

[10] Jian, Y., Pasquier, M., et al., Healthcare,Vol 9(12).

[11] Alex, S.A., et al. Deep convolutional neural network for diabetes mellitus prediction, 2022.

[12] Fayaz, R., Reddy,et al., An Intelligent Harris Hawks Optimization (IHHO) based Pivotal Decision Tree (PDT) Machine Learning Model for Diabetes Prediction

Dutta, A. et al. (2022) discussed that a significant drawback is the lack of considered information and the presence of outliers and missing information in clinical datasets, making this a challenging endeavor. To resolve the issue, an ensemble of ML Methodologies was deployed. The deployed methodology is essential in certifying reliable and exact predictions, allowing this study to accomplish its target of initial diabetes prediction.

Ahmad, H. F et al. (2021) discussed how premature stoppage could restrain difficulties and influence a person's quality of life, lowering costs and positively impacting communities and the healthcare system. To attain the objective, an ML classifier was deployed. The deployed methodology was applied to the information to recognize hazard aspects and their subsidiary influence on diabetes sorting.

Gadekallu, T.Ret al. (2020) carried out the GWO methodology to categorize the take-out characters of a diabetic dataset. GWO enables the choice of ideal factors for training the DNN method. The deployed methodology's main drawback was its low dataset classification performance.

Gupta, H. et al. (2022) discussed the main aim of developing a predictive device based on a diabetes dataset to assist health professionals in reducing diabetes-related mortality. A Quantum ML (QML) methodology was deployed to attain this objective. The proposed method has a low accuracy rate for predicting diabetic datasets.

Sharma, A et al. (2021) developed a supervised ML methodology for predicting early-stage diabetes. These programs make assumptions based on various factors, including health history and lifestyle. They learn from multiple examples of diabetics and non-diabetics to make more accurate guesses.

Rajkumar et al. (2024) described algorithms proposed low efficacy, which can cause a delay in diagnosis and make the process of developing a personalized treatment plan more difficult. On the other hand, early identification of breast cancer can affect the diagnosis's accuracy and be expensive.

---

[13] Dutta, A., Hasan, M. K.,et al., International Journal of Environmental Research and Public Health, Vol 19(19).
[14] Ahmad, H. F., et al., Applied Sciences, 11(3), 1173. https://doi.org/10.3390/app11031173.
[15] Gadekallu, T.R., Khare, N., et al., Deep neural networks to predict diabetic retinopathy, 2023.
[16] Gupta, H., Varshney, et al., Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction, 2022.
[17] Sharma, A., Guleria, K., et al., Prediction of Diabetes Disease Using Machine Learning Model. vol 733. Springer.
[18] Rajkumar, R., Gopalakrishnan, S.,et al., Mesopotamian Journal of Artificial Intelligence in Healthcare, 2024.

U. Ahmed et al. (2022) discussed the utmost importance of early diagnosis and prevention in medicine. Knowing an illness's indications is imperative for foreseeing it. To attain this objective, an ANN and SVM methodology were deployed. The outputs of these models become input membership functions of the fuzzy model, which, ultimately, governs whether the diabetes analysis is negative or positive through fuzzy logic.

Wee, B.F et al. (2024) discussed that most researchers use datasets recording laboratory-based clinical measurements for model training and validation processes. However, these predictive models are considered overestimated. Measurements based on lab tests are cast-off to predict and analyze whether someone has diabetes and can be expected to be accurate. An ML and DL protocol was deployed to attain the objective protocol. This approach suggests a database of non-invasive laboratory measurements. Nevertheless, it is still essential to analyze subject characteristics and diabetes outcomes in advance to avoid prohibitive costs associated with database construction.

Olisah, C. Cet al. (2022) discussed that if not adequately managed or misdiagnosed, diabetes can threaten vital organs like the eyes, kidneys, and nerves. However, judging from the current accuracy, much room still exists to improve. To improve accuracy, a twice-growth DNN (2GDNN) methodology was deployed. The proposed method does not yet address some of the limitations of the diabetes dataset, as only information on adult male and female patients is provided. Therefore, there is a need for a form of expression for men, women, and children.

Nadeem, M. Wet al. (2021) discussed that developing data-driven functions and facilities for diagnosing and classifying critical medical conditions is challenging due to the low volume and worsened circumstantial information for algorithm training and validation, resulting in a loss of accuracy. A fusion ML methodology was deployed to resolve the issue. The proposed method is used for diabetes identification and foreseeing the beginning of severe proceedings in diabetic patients.

Abdüssamed Erciyaset al. (2021) carried out Faster RCNN based on DL methodology to automatically and self-sufficiently identify datasets and determine how lesions are classified. In this deployed methodology, lesions are examined, and the region of interest is noticeable. The outcomes show that the deployed methodology obtained unsuccessful results.

---

[21] Wee, B.F., et al. Diabetes detection based on machine learning and deep learning approaches.
[22] Olisah, C. C., Smith, L., et al., Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Computer Methods and Programs in Biomedicine, Vol 220.
[23] Nadeem, M. W., et al., Healthcare, 9(10), 1393. https://doi.org/10.3390/healthcare9101393
[24] Abdüssamed Erciyas, et al., "An Effective Method for Detecting and Classifying Diabetic Retinopathy Lesions Based on Deep Learning.

A. C. Lyngdohet al. (2021) discussed that the foremost objective is to find the optimal results, precision, and computational time for foreseeing diabetes disease. To attain this objective, a novel ML protocol was deployed. However, the deployed protocols cannot produce stable and good accuracy when visualizing the training testing accuracy and checking the model overfitting and underfitting.

Sheela, M et al. (2024) discussed Although the fuzzy logic method has been applied to the classification of lung illness prediction, it has encountered several obstacles, including inaccuracies in the output and problems identifying segmented sections.

S. K Net al. (2022) carried out a Latent Dirichlet Allocation collated with ANN (LDA-ANN), which was deployed to execute an operative diabetes classification. LDA-ANN faces problems in accurately predicting diabetes in patients. Additionally, the classification accuracy was weakened when assessed for massive datasets.

A. Jabbar et al. (2024) used an Adaptive Particle Swarm Optimizer (APSO) methodology to classify the severity levels of diabetic retinopathy datasets. However, previous studies have addressed the DR classification problem through retinal fundus datasets; the proposed methodologies detect isolated lesions but detect isolated lesions, but they are unable to identify all lesions simultaneously and need a comprehensive framework.

Dutta, R.R. et al. (2024) discussed that its purpose is to collect raw data, preprocess the data to remove anomalies in the obesity dataset, and transform that data into a format that ML methods can more easily and efficiently process. It includes data preparation, such as aggregation, reduction, and refinement. This work demonstrates various visualizations using a dataset of obesity traits.

Yu, J. et al. (2024) discussed developing an ASCVD risk prediction methodology that integrates long-term hazard features via DL. To attain the object, a Dynamic-Deep Hitmethodology was deployed. Uniting longitudinal hazard features into ASCVD hazard prediction via DL cannot get better methodology identification and standardization.

---

[25] A. C. Lyngdohet al., "Diabetes Disease Prediction Using Machine Learning Algorithms.
[26] Sheela, M., et al., Machine learning based Lung Disease Prediction Using Convolutional Neural Network Algorithm.
[27] S. . K N and R. P. . K N, “Diabetes Mellitus Disease Prediction and Classification using Latent Dirichlet Allocation and Artificial Neural Network Classifier.
[28] A. Jabbar et al., "A Lesion-Based Diabetic Retinopathy Detection Through Hybrid Deep Learning Model.
[29]Dutta, R.R., et al., Obesity disease risk prediction using machine learning,2024.
[30]Yu, J., Yang, X., et al., Incorporating longitudinal history of risk factors into atherosclerotic cardiovascular disease risk prediction using deep learning. Scientific Reports, Vol 14(1).

While most algorithms have been used to predict diabetes, they are inaccurate in assessing accuracy. However, effective computational models are required to determine the severity of te various risk classifications associated with diabetes. Existing systems cannot determine accuracy. As a result, optimization approaches are required to improve performance metrics. High computing time in existing systems is a significant issue that must be addressed to reduce calculation time and deliver accurate predictive models. Overcome the inadequacies of precision assays and develop effective approaches for predicting the severity of various diabetes risk factors. Proposed a diabetes prediction technique based on RNNRBFN that can learn more complicated patterns and enhance classification performance over utilizing the network in isolation.

## III. IMPLEMENTATION OF PROPOSED METHOD

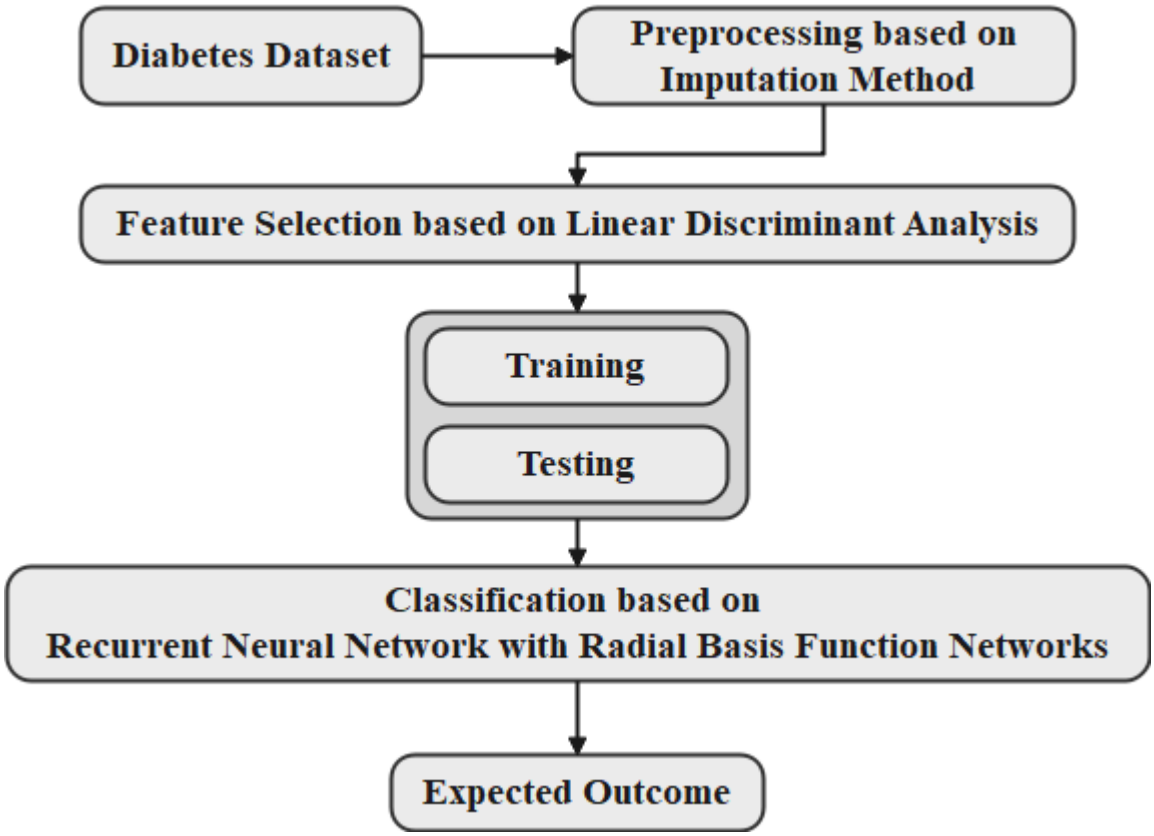


**Fig. 1 Proposed Methodology's Architecture Diagram**

The details of the methodology used are indicated in Figure 1. In this paper, the diabetic dataset is obtained from the Mendeley Data website.

The imputation method is used to exchange the absent characteristic in the dataset with an alternative characteristic so that most of the data or info in the dataset is retained. After

preprocessing, the LDA methodology is used for feature selection; it is used to find a set of linear discriminants that maximizes the proportion of between-class characters to within-class characters. The proposed method is used to learn more complex patterns and improve classification performance compared to using either network in isolation.

## A. Dataset Collection

The diabetes dataset from the Mendeley Data Website was used in this study. The dataset utilised in this work is comprised of 1000 patient records that have 12 specified criteria. Gender, age, urea, creatinine, HbA1C, cholesterol, triglycerides, HDL, LDL, VLDL, BMI, and class are some of these traits. The patients' pre-diabetic, diabetic, and non-diabetic statuses are indicated by these goal categories. The diabetes dataset's distinctive types are detailed in depth in Figure2.

**Fig 2. Details of characteristics types**

| S.No | Characteristics | Parameters Description |
|---|---|---|
| 1 | Gender | Male or Female |
| 2 | Age | In years, min-max value (20-79) |
| 3 | Urea | In mg/dl, min-max value (0.5-38.9) |
| 4 | Cr | In µmol/L, min-max value (48,80) |
| 5 | HbA1c | In mmol/L min-max value (0.9-16) |
| 6 | Chol | In mmol/L min-max value (0.0-10.3) |
| 7 | TG | In mmol/L min-max value (0.3-13.8) |
| 8 | HDL | In mmol/L min-max value (0.2-9.9) |
| 9 | LDL | In mmol/L min-max value (0.3-9.9) |
| 10 | VLDL | In mmol/L min-max value (0.1-35) |
| 11 | BMI | min-max value (19-47) |
| 12 | Class | Pre-Diabetic (Y), Diabetic (P), and Non-Diabetic (N) |

## B. Imputation Method

In this part, we preprocess the diabetes dataset using the Imputation method. This method allows us to replace missing data with alternative characteristics, ensuring that most of the dataset's information is retained. We use this method because removing data from the dataset each time is impractical, which can significantly reduce its size. We implement three simple missing value methodologies: constant, mean value attribute, and unspecified characteristic value replacement. The performance of these methodologies is thoroughly assessed for different fractions of missing values in the dataset using various clustering techniques. The diabetic dataset is used as the basis for this comprehensive assessment.

Missing values can occur in conditional or class properties. To handle missing values, for example, we can ignore characteristics with missing values, fill in missing characteristics manually, switch missing characteristics with a comprehensive invariable or the median of the characteristics, or often fill methods with a value taker. Non-parametric imputation methodology can afford a better fit by getting the constitution in the data set (note that it may only provide a better fit if the parametric model is correctly specified).

## C. Handle Missing Values

Choosing an appropriate technique depends on the problematic domain, the data domain, and the data preprocessing goals. There are several ways to carry out missing characteristics in a dataset, including:

### Disregard the data row:

This is habitually executed as soon as the class categorization is missing or when there are too many attributes from the array. However, a high proportion of such arrays degrades the performance. For example, suppose a dataset of diabetes registry data (age, sex, urea, etc.) and a column that sorts the values from lowest to highest. Suppose we aim to develop a methodology that forecasts whether a person has diabetes. Data rows with columns with missing values cannot be used to predict diabetes and are often disregarded and deleted in advance of proceeding with the method.

### Castoff a global constant to fill in for missing values:

To load in the missing characteristics, determine a novel global constant characteristic, such as "Unknown," "N/A," "Hyphen," or Infinity. The reason for using this technique is that trying to predict missing values may need to be clarified. For example, we have a diabetes dataset, and some patient details do not have resident attribute data. Compared to using something like "N/A," going to a level does not make sense.

### Using of Attribute Mean:

This method replaces missing characteristics for a feature with the median of the attribute in the data set. In a data set with diabetes threshold values, if the mean of diabetic patients is X,

then that value can be used to replace the missing diabetes value.

## D. Missing Value of Imputation Calculation Method

Validating clustering methodologies and comparing the performance of different algorithms is complicated because finding objective measures of clustering quality is difficult. Comparing results to external standards requires a degree of consistency. Since we assume that each record is assigned to only one type of external criterion and assigned to only one cluster, we can apply consistency between the two partitions.

Assumed a $n$ of characteristics $P = \{C_1, \dots, C_0\}$ and the cluster of two data's of $P$which we need to relate $A = \{A_1, \dots, A_I\}$ and $B = \{B_1, \dots, B_P\}$ where the dissimilar subsets of $A$ and $B$are divide and their union is equivalent to P, and then we calculate the ensuingassessments:

$$o = m \times \frac{(x+y)+k}{(x+y+i+z)+p} = m \times \frac{(x+y)+k}{\binom{n}{2}+p} \quad (1)$$

let, $x$ is the integer of characteristics in $P$ that are in the identicalattribute as $A$ and in the similar partition as $B$. $y$ is the integer of characteristics in $P$ that are neither in the identical partition as $A$ nor in the similar partition as $B$. $i$is the integer of characteristics in $P$ which are in the identical partition as $A$ and not in the similar partition as $B$. $z$ is the integer of characteristics in $P$ which arenot in the identical partition as $A$but are in the similar partition as $B$.Instinctively, $x + y$ as the number of matches between $A$ and $B$ and $i + j$ as the number of discrepancies between $A$ and $B$.

Let S be a data set of R records, each record containing T characteristics. Hence, this data set contains RxT characteristic values. If dataset S has a missing characteristic value, it is epitomized by an innumerable string in the dataset. Here, we present the algorithmic steps to impute missing values in such dataset S.

### Algorithm Step:

Phase 1: for $p = 1\ toT$
    for $i = 1\ toR$
        if $S(p, i)$ is not a number then alternate zero to $S(p, i)$
Phase 2: for $i = 1\ toR$
    find accuratenesscharacterArof all the characteristics of the column $i$
    $Ar(i)$ = (sum of all values of column i of s)/n
    for $p = 1\ toT$
        for $i = 1\ toR$
            if $S(RxT)$ is a missing value then alternate $Ar(i)$ to $S(RxT)$
Phase 3: for $i = 1\ toR$
    find the accurateness character Ar of all the characteristics of the column $i$

Min(i) = (minimum of all the characteristics of i)

Max(i) = (maximum of all the characteristics of i)

for $p = 1\ toT$

for $i = 1\ toR$

if S (T, R) is a missing value, then alternate an undefined value among Min(i) and Max(i) to S (T, R)

When preprocessing diabetes datasets for disease prediction, imputation methods are important in handling missing values. A common strategy uses the mean, median, or mode of the pertinent feature to assign missing qualities. The type of data set and the amount of missing data should determine the best imputation technique. It is essential to estimate the carrying out of different imputation techniques using metrics, for instance accuracy, F1 score, etc.

## E. Linear Discriminant Analysis (LDA) Method

After the preprocessing, the dataset is ready to select the features based on the LDA method. LDA is used to reduce variables and efficiently detect diseases. For this diabetes dataset, characters were circulated to LDA, and performance was analyzed using attribute reduction. Here, characters are condensed into four variables, and these certain variables provide respectable accurateness. By keeping as much category information as possible, LDA is used to reduce proportion by maximizing its statistical qualities with the data that is provided. Gaussian covariance is calculated for several variables, while the mean and variance are calculated for a single variable input. This study's LDA technique is displayed below.

### Algorithm Step:

Phase 1: Load the S (T, R)

Phase 2: Determine the mean vector by finding the mean for each attribute in the dataset that corresponds to a distinct class.

Phase 3: Compute the information contained in the class scatter matrix, then save the attributes in Sb.

$$S_b = \sum_{j\in\omega_b}(j - d_b)(j - d_b)^T \quad (2)$$

Phase 4: Compute the information contained in the class scatter matrix, then save the attributes in Sj.

$$S_j = \sum_{b=1}^{e} E_b(D_b - d)(D_b - d)^T \quad (3)$$

Phase 5: Multiply the inverse of Sb by Sj b castoff to find the Eigen vector, which is given by the equation Xv=λv, where v is the Eigen vector and **s** is the Eigen value.

$$X = S_b^{-1}S_j \quad (4)$$

Phase 6: Save the values in descending order to a new array variable. Vectors with the lowest eigenvalues are omitted.

Phase 7: Repeat phase 3 until every feature has been chosen.

Using the LDA approach for feature selection to predict diabetes risk from diabetes datasets has a number of benefits. With the retention of discriminative information, LDA efficiently identifies the most pertinent features and minimizes dimensionality. This could boost the efficacy and accuracy of diabetes risk prediction models as well as their performance. LDA also aids in the interpretation of the prediction outcomes and offers insights into the underlying data structure. But in order to guarantee the model's generalizability and dependability, it's crucial to thoroughly assess the robustness and relevance of the chosen characteristics across several datasets.

## F. Recurrent Neural Networks with Radial Basis Function Networks (RNN-RBFN)

In this section, the diabetes dataset is classified by the use of the RNN-RBFN method. RNNs are suitable for processing continuous data and modeling the temporal dependence of diabetes risk factors over time. On the other hand, RBFN is known for its ability to approximate complex nonlinear functions, which helps capture complex relationships among various risk factors. Our goal is to improve each of these two neural network designs individually in order to increase the precision and resilience of diabetes risk prediction models.

In the context of diabetic risk prediction using diabetes datasets, RBFNs can be used as a classification method. The diabetes dataset containing information relevant to diabetic risk is preprocessed, which may include steps like normalization, splitting into training and testing sets, and feature selection. The RBFN is trained in feature selection. The network gains knowledge of the connection between input characters—such as patient data—and the intended class through training. These functions calculate the separation in feature space between a data point and the centre point (max-min value of all the features). Once trained, the RBFN calculates the weight of new data points into different diabetic risk levels based on the learned characteristics from the training dataset. RBFN consists of two different layers of training. The hidden layer consists of RBFN, and the output layer consists of linear weights and sigmoid functions. Because these layers behave differently, the training activity is divided into two parts.

## G. Hidden Layer

Nodes in this layer can use specific parameters based on different kernel capabilities. In this part, we use a simplified Gaussian function (6). This function has two internal parameters: the center position (c) and the coverage area (β), which are cast for training intents. We executed the k-means methodology to choose the preliminary state of each RBF. Following are the training steps:

$$g(l) = xe^{\frac{-\|l-y\|^2}{2i^2}} + z \quad (6)$$

This formula has four internal parameters as $x, y, i, andz$ and one input parameter as $l$.

This layer can define different Gaussian functions by changing the parameter usage while

maintaining the structure of this equation.

- Bring together same label training data composed,
- Set "y" as centre position of all training data,
- Estimate average distance "d" by using all points,
- Set $\beta = (2\,d^2) - 1$ (7)
- Weight is 1 for classified outputs, otherwise 0,

## H. Output Layer

Nodes in this layer contain weights $m_x$ and thresholds for each RBFN node to produce specific results. The main objective of RBFN is to determine if a given input belongs to one or more known classes. Therefore, the nodes in this layer should consider the binary output.

$$Output = \sum m_x RBF_x > Threshold \quad (5)$$

Since the training data output is already provided to the patient record for training purposes, the weights are increased or decreased until the correct output value is reached. From the definition and training method of RBFN, it is clear that RBFN has a much simpler structure as it has only three layers and is very flexible based on its ability to grow in size. With this simplified function, the function only has two internal parameters to change the state of the function and the region it covers. The Recurrent Neural Networks (RNN) are introduced to improve the classification of the dataset.

## I. Improving Classification

The RNN was deployed to enhance the classification of the weight module. As the proposed method is designed to handle sequence data, it is necessary to structure the dataset properly. It involves generating a sequence of data points. Each series represents the patient's data over time. For diabetes risk prediction, the RNN unit LSTM is used. The input to the method is a sequence of characters (e.g., CR, Urea, etc.), and the output is the predicted risk of diabetes (e.g., diabetic, non-diabetic, prediabetic). The RNN learns how to map the input sequences to the corresponding diabetes risk characters. The input sequence is routed through the network, the loss is computed, and the network weights are gradually updated by backpropagation. Utilize the validation set to assess the model's performance following training, and modify the hyperparameters to avoid overfitting. Lastly, a test set is used to evaluate the model's generalization capacity. We utilize RNN as a hidden layer implementation unit to method the semantic representation of characteristics. Generally, each unit of RNN is calculated as follows.

$$A = \begin{bmatrix} j_{t-1} \\ A_t \end{bmatrix} \quad (8)$$

$$g_t = \sigma(K_g.A + y_g)$$ (9)

$$m_t = \sigma(K_m.A + y_m)$$ (10)

$$n_t = \sigma(K_n.A + y_n)$$ (11)

$$p_t = g_t \odot p_{t-1} + m_t \odot \tanh(K_p.A + y_p)$$ (12)

$$r_t = m_t \odot \tanh(n_t)$$ (13)

$K_m, K_g, K_n \in \times ifo ,, \mathbb{R}^{dx2d}$ is the weight matrix, $y_g, y_m, y_n \in \mathbb{R}^d$ are the partialities of the LSTM learned all through the training, parameterization and transformation process of the input gate, disremember gate and output gate, correspondingly. $\sigma$ is the sigmoid process and $\odot$ is the character-wise multiplicity. $j_t$is the input of the LSTM unit cell, and ht is the hidden state at time $t$.

Suppose a dataset consists of Cnumbers, each number$l_x$ containing $T_x$numeric. $t \in [1, T]$ with wit denotes the word in the $x$number. Intended forcharacter computations, $i_t$signifies the numeric vector $v_t$implanting. The first hidden layer $o_{xt}$vector $t \in [1, T]$ and is castoff to characterize the number. For numeric-level computation, $x_t$epitomizes the number embedding vector $l_x$ . In this situation, the hidden layer vector $o_x x \in [1, C]$is used to denote the dataset.

Consider the hidden layer to be a representation of the dataset, and place a softmax layer on top of it to predict the class label. {Diabetic, Non-Diabetic} or {Pre-Diabetic} reports pertaining to diabetes. More importantly, the output layer of the RNN considers two likely labels. Therefore, we developed two discrete methodologies to categorize diabetic or not and pre-diabetic. Considering $o^*$ as the final representation of the diabetes values, the softmax layer can be built as:

$$b = softmax(K_l o^* + y_l)$$ (14)

here, $K_x$ and $y_x$are components of the layer called softmax. As the training loss function, we take use of the unfavorable log-likelihood of the exact labels.

$$C = -\sum_d log b_{dz}$$ (15)

$z$ is the label of document $d$,

In an RNN system, the hidden circumstances are nourished to anormal pooling layer to acquire the numericillustration and the lastbinarydepiction. Perhaps, the last feature depiction of a radiographic report may be calculated as:

$$o^* = \sum_{x \in [1,C]}^{x} o_i$$ (16)

We deploy a domain phrase attention protocol to detention the maximum imperative parts of the dataset given number-level domain expressions. As shown, it makes sense to recompense characteristics that provide hints for appropriately classifying datasets. Therefore, pay special attention if domain phrases are in the numeric. We scramblerespectively domain term as a randomly initialized unceasing real-valued vector $q\varepsilon\mathbb{R}^d$.

$$e_x = \tanh(K_l o_x + K_{dq} q + y_l) \quad (17)$$

$$\alpha_x = \frac{\exp(e_x^T e_l)}{\sum_x exp(e_x^T e_l)} \quad (18)$$

$$o^* = \sum_{x\in[1,C]}^{x} \alpha_x o_x \quad (19)$$

$K_l \& K_{dq}$ is known for projection parameters and $y_l$ is represent for bias parameter.

It is especially helpful to employ RNN for diabetes dataset classification when working with continuous or time-dependent data. The proposed method is well suited for this task because it can capture patterns and dependencies in data over time. We can choose to pre-process the data set, including normalizing the data and reshaping it into an array, to certify that it is in a layout appropriate for the RNN. Next, build an RNN-RBFN method that learns from recurring patterns in the data and classifies diabetes risk levels. Figure 3 expressions a flowchart of the deployed methodology that visually shows all the steps performed to arrive at the RNN-RBFN technique.

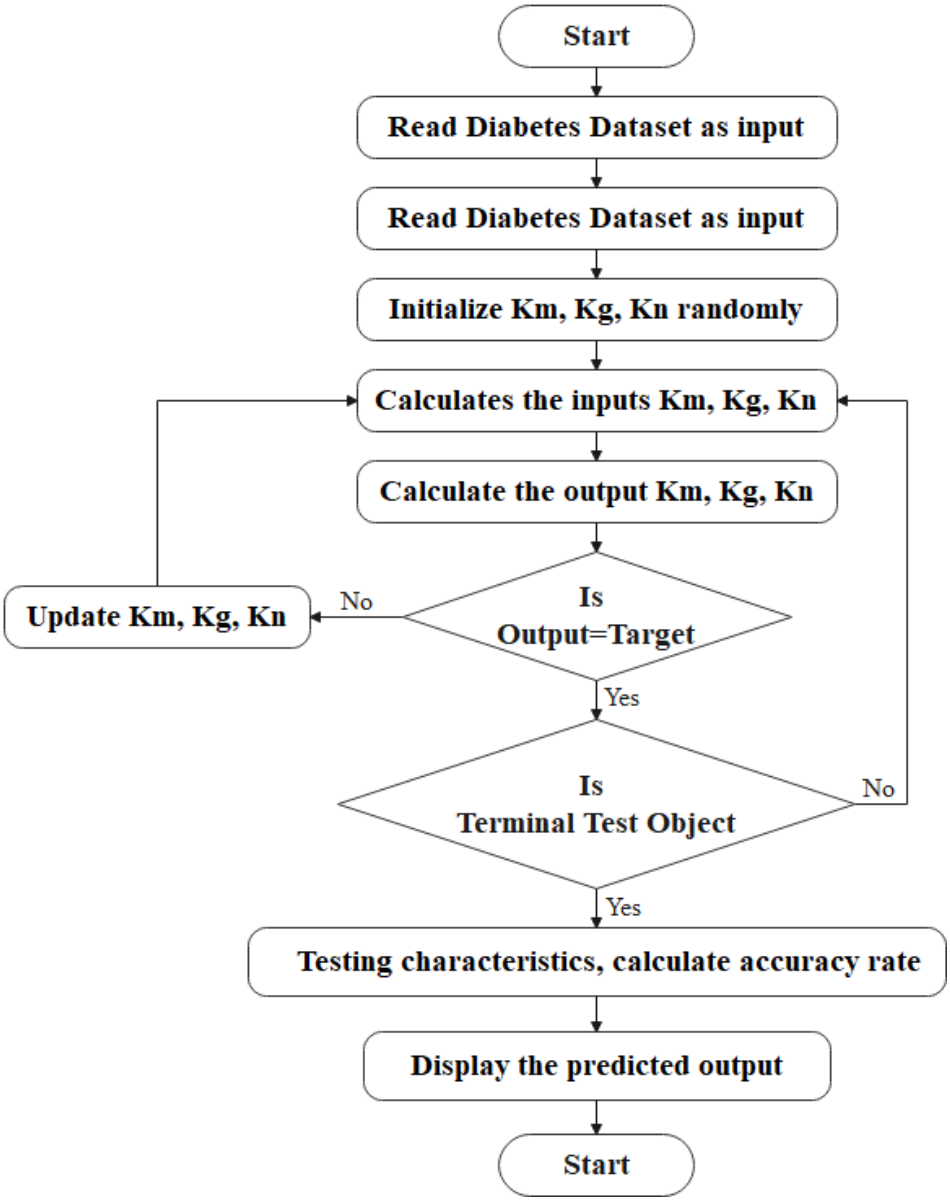


**Fig.3. Flowchart Diagram for RNN-RBFN method**

RNNs can handle sequential data efficiently and are valid for time series data sets such as medical records. RBFN is known for its ability to approximate complex functions and can enhance the performance of RNNs by providing a more efficient representation of data sets. This combination may improve the accurateness and effectiveness of diabetes risk prediction when compared to RBFN.

## IV. RESULT & DISCUSSION

The execution of the introduced methods is estimated using precision, recall, accuracy, F1 score and time complexity. This evaluation uses the existing methods RT, LDA-ANN and APSO to find reliable and accurate diabetes detection methods with the deployed methodology

.

**Table 1.Parameters for Simulation**

| Parameters | Values |
|---|---|
| **The dataset's name** | Diabetes Dataset |
| **No. of Records** | 1000 |
| **Language** | Python |
| **Tool** | Anaconda |

Table 1 demonstrations the simulation consequences and the parameters of this paper. The diabetes dataset is used to classify the patient's medical history and consists of 1000 datasets. The implementation method is to use Python and the Anaconda tool to implement it.

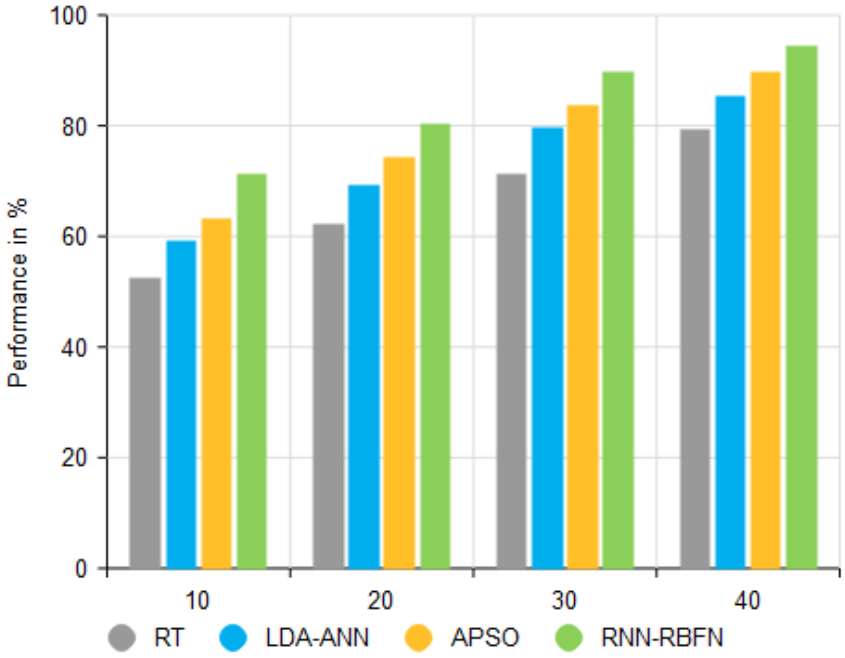


**Fig.4. Precision performance analysis in %**

The figure 4 expressions that the precision execution of the RT is 79.35%, LDA-ANN is 85.47%, APSO is 89.63% and the precision performance of the RNN-RBFN is 94.35%. The deployed methodology has the better precision than the previous methodology. Because the

employed technology successfully captures temporal dynamics and complicated interactions in the data, it can predict diabetes risk with more precision and accuracy.

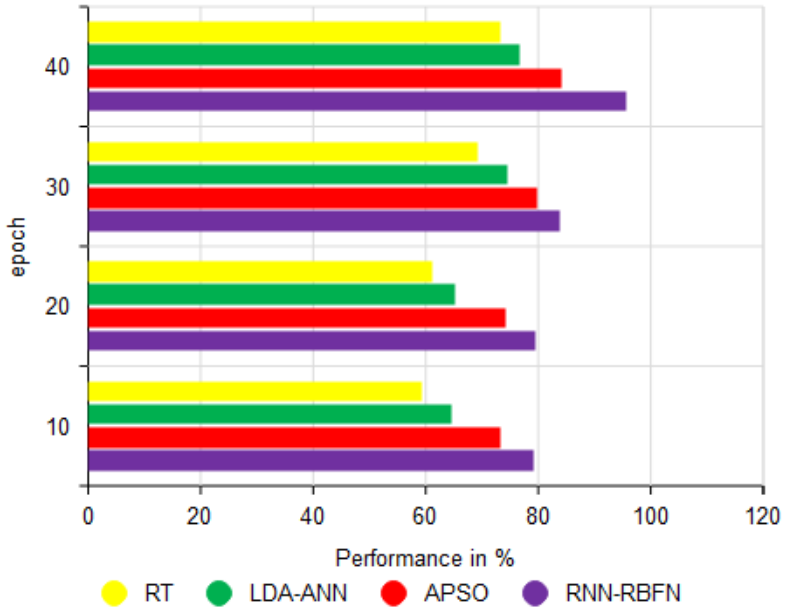


**Fig.5. Recall performance analysis in %**

The figure 5 expressions that the recall execution of the RT is 73.21%, LDA-ANN is 76.54%, APSO is 84.21% and the recall performance of the RNN-RBFN is 95.62%. The deployed methodology has the better precision than the previous methodology.The high recall of the RNN-RBFN method for diabetes risk prediction suggests that the model effectively identifies most of the true positive events from the dataset. This is beneficial because the model can identify individuals at true risk with sensitivity, thus reducing the potential for false positives.

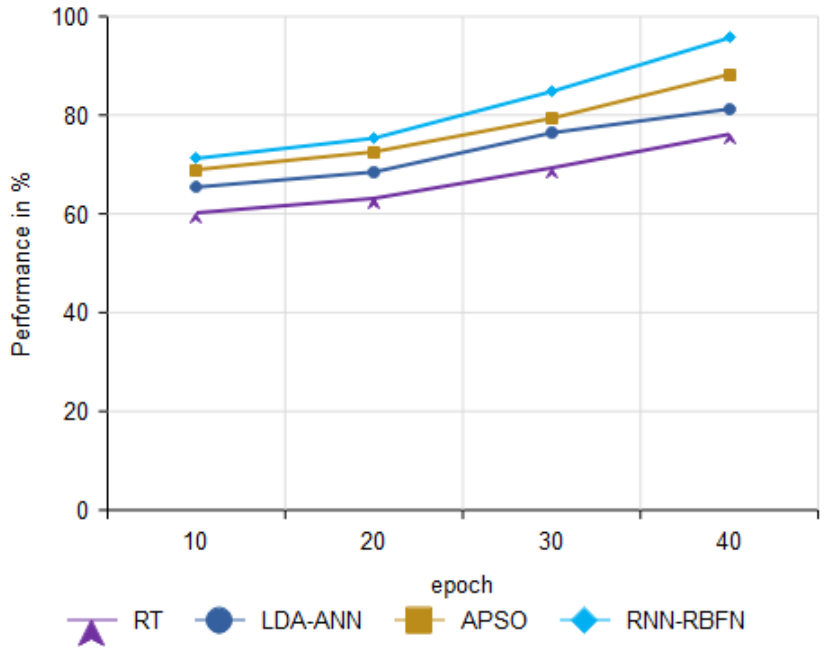


**Fig.6. Accuracy performance analysis in %**

The figure 6 expressions that the accuracy execution of the RT is 76.12%, LDA-ANN is 81.23%, APSO is 88.24% and the accuracy performance of the RNN-RBFN is 95.64%. The deployed methodology has the better accuracy than the previous methodology. The high accuracy values of the RNN-RBFN method for diabetes risk prediction indicate that it performs generally well in accurately classifying positive and negative events from the dataset. This is advantageous because the model produces less prediction errors and provides reliable results.

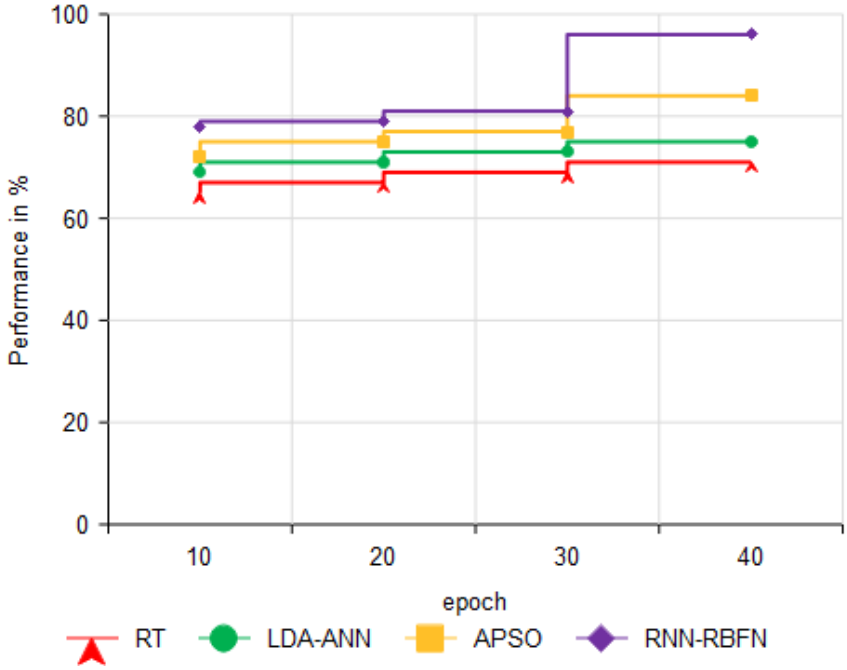

**Fig.7. F1 score performance analysis in %**

The figure 7 expressions that the F1 score execution of the RT is 71%, LDA-ANN is 75%, APSO is 84% and the accuracy performance of the RNN-RBFN is 96%. The deployed methodology has the better F1 score than the previous methodology.A prominent F1 score specifies that the method has togetherlarge level of precision and large level of recall, which allows it to effectively and accurately detect people with diabetes while reducing false positives, meaning it exhibits excellent performance.

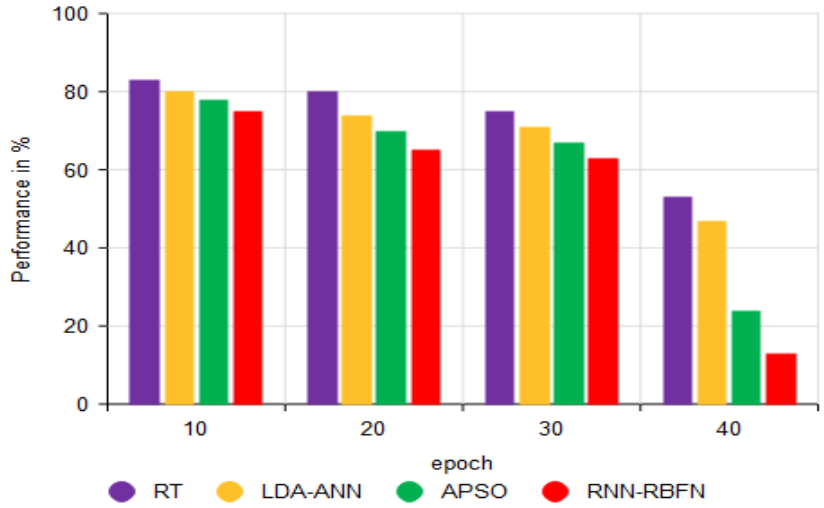


**Fig.8. Time complexity performance analysis in %**

The figure 8 illustrations that the time complexity performance of the RT is 53%, LDA-ANN is 47%, APSO is 24% and the time complexity performance of the RNN-RBFN is 13 %. The deployed methodology the time complexity is very low than the previous methodology.The low time complexity of the RNN-RBFN method in diabetes risk prediction means that the model can predict rapidly and efficiently. This is beneficial because it makes it possible to handle huge data sets quickly, which is crucial in clinical settings where prompt predictions can lead to prompt action and better patient outcomes.

## V. CONCLUSION

Our research paper deployed an RNN-RFBN method for the diabetic dataset classification. Similarly, this paper usedthe diabetes dataset to evaluate a narrative nature and comprehensive of the Imputation method. We used Imputation method is castoff to exchange the missing data in the dataset with alternative characteristics so that furthermost of the data or info in the dataset is retained. This method is used because it is impractical to remove data from the dataset every time and can significantly reduce its size. Then the preprocessed dataset is selecting the features with the use of LDA method. RNN-RBFN can potentially learn more

complex patterns and improve classification performance compared to using either network in isolation. Based on these variables, the method is taught to forecast an individual's risk of acquiring diabetes, offering insightful information for early intervention and customized treatment plans. Moreover, the proposed method performed well and accomplished a precision of 94.35%, recall 95.62%, accuracy 95.64%, F1 score 96%, and time complexity 13% using the diabetes dataset. Better techniques might facilitate early detection and action by medical professionals, which would eventually benefit patients and save expenditures.