Hybrid RNN-RBFN Model for Accurate Diabetes Prediction Using Evolutionary Gravitational Search and Dynamic Incremental Normalization

[1]Baseera A, [2]Dhiyanesh B, [3*] Parveen Begam Abdul Kareem, [4]Shanmugaraja P, [5]Anusuya V, [6]Shermy R P
[1]VIT Bhopal University, Bhopal, India
[2] SRM Institute of Science and Technology, Vadapalani Campus, Chennai.
[3*] Taibah University, Yanbu, Saudi Arabia.
[4]Sona College of Technology, Salem, Tamil Nadu, India.
[5]Ramco Institute of Technology, Rajapalayam, Tamil Nadu, India.
[6] Dr. N.G.P. Institute of Technology, Coimbatore, Tamilnadu India.
[1]basheera.zain@gmail.com, [2]dhiyanu87@gmail.com, [3*] parveenkareem@gmail.com,
[4]shanmugarajap@gmail.com, [5]pgkrishanu@gmail.com, [6]shermyraj103@gmail.com

Abstract

Diabetes is a chronic metabolic illness marked by high blood glucose levels, which can lead to serious complications if not recognized and treated promptly. Diabetes must be predicted accurately and early to allow for appropriate intervention and treatment planning. This study seeks to create a high-performance predictive model for diabetes diagnosis by combining advanced preprocessing, feature selection, and classification techniques. A unique hybrid strategy is proposed, which uses Dynamic Incremental Normalization to effectively normalize the data distribution and the Evolutionary Gravitational Search Algorithm to discover the most significant features, improving model accuracy and efficiency. A hybrid Recurrent Neural Network (RNN) and Radial Basis Function Network (RBFN) model is used for classification, taking use of RNN's capacity to capture sequential dependencies and RBFN's ability to handle nonlinear decision boundaries. The model was tested with the PIMA Indian Diabetes Dataset and 10-fold cross-validation, yielding an accuracy of 99.5%, precision of 98.9%, recall of 99.1%, and F1-score of 99.3%. **Keywords:** Diabetes, Neural Network, Evolutionary Gravitational, Normalization, PIMA Indian Diabetes Dataset, Clinical Decision Support Systems.

1) Introduction

The rising incidence of diabetes in India presents a serious healthcare concern. A startling 82 million persons in India currently have diabetes, according to data from the International Diabetes Federation. The estimates that this count will rise to a significant number of 128 million by 2040 are even more concerning. This means that about one in seven Indian adults may be at risk for diabetes [1]. Progressive urban expansion in India is associated with modifications to lifestyle that increase the risk of diabetes. These shifts include increasing numbers of obese people, a higher intake of refined foods, and decreased levels of vigorous exercise. Furthermore, external factors, anxiety, and hereditary traits all come into influence [2].

----

[1] J.J. Khanam, S.Y. Foo, A comparison of machine learning algorithms for diabetes prediction, ICT Express 7 (2021) 432–439, http://dx.doi.org/10.1016/J.ICTE. 2021.02.004.

[2] V. Jaiswal, A. Negi, T. Pal, A review on current advances in machine learning based diabetes prediction, Prim. Care Diabetes 15 (2021) 435–443, http://dx.doi.org/10.1016/j.pcd.2021.02.005.

Tragically, a large number of diabetes cases go undetected until problems develop, underscoring the necessity of prompt identification techniques and greater awareness. With diabetes as the seventh highest contributor to mortality in India, the illness has a huge humanitarian impact [3]. The financial cost is similarly worrisome. According to estimates, diabetes costs the Indian economy nearly one hundred billion dollars a year. It is obvious that tackling this escalating worldwide health emergency calls for a multifaceted strategy that emphasizes type 2 diabetes prevention, timely identification of all forms of the disease, and appropriate treatment of the condition [4]. Given these difficulties, there is a pressing need for sophisticated medical devices that can precisely and effectively identify diabetes.

Although helpful, conventional methods for diagnosis are limited in both precision and sensitivity. This has prompted scientists to investigate novel strategies, such as genomics analytics. The use of microarray data on gene expression to find indicators for Type II Diabetes Mellitus has showed potential [5]. Microarray data's great dimensionality, nevertheless, comes with its own variety of difficulties. In machine learning models, overfitting and decreased applicability may result from the large number of genomes in comparison to the often small number of samples [6]. To find the most pertinent genes for the diagnosis of diabetes, this "curse of dimensionality" calls for efficient extraction of features and choice of methods. Deep learning methods have been popular in healthcare investigations in the past few decades, with the promise of more precise and effective illness identification.

In this work, metaheuristic algorithms are employed, which are agnostic of problems and use fitness functions to direct their quest for ideal solutions [7].

----

[3] M.E. Hossain, S. Uddin, A. Khan, Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes, Expert Syst. Appl. 164 (2021) 113918, http://dx.doi.org/10.1016/j.eswa.2020.113918.

[4] M.R. Islam, S. Banik, K.N. Rahman, M.M. Rahman, A comparative approach to alleviating the prevalence of diabetes mellitus using machine learning, Comput. Methods Programs Biomed. Update 4 (2023) 100113, http://dx.doi.org/10.1016/J.CMPBUP.2023.100113.

[5] M. Bernardini, M. Morettini, L. Romeo, E. Frontoni, L. Burattini, Early temporal prediction of type 2 diabetes risk condition from a general practitioner electronic health record: A multiple instance boosting approach, Artif. Intell. Med. 105 (2020) 101847, http://dx.doi.org/10.1016/j.artmed.2020.101847.

[6] M.M.F. Islam, R. Ferdousi, S. Rahman, H.Y. Bushra, Likelihood Prediction of Diabetes at Early Stage using Data Mining Techniques, Springer, Singapore, 2020, http://dx.doi.org/10.1007/978-981-13-8798-2_12.

[7] D. Parkhi, N. Periyathambi, Y.G. Weldeselassie, V. Patel, N. Sukumar, R. Siddharthan, L. Narlikar, P. Saravanan, Prediction of postpartum prediabetes by machine learning methods in women with gestational diabetes mellitus, iScience 26 (2023) 107846, http://dx.doi.org/10.1016/J.ISCI.2023.107846.

Interestingly, there is no established procedure for choosing the optimal parameters for deep learning classification algorithms, which prevents them from improving their performance based on recognized proof and empirical trends. The suggested distributed metaheuristic optimization technique shows promise in filling this gap by improving resilient effectiveness for diabetes classification challenges on unexplored datasets. The investigation of diabetic data balance and classifier parameter tweaking using metaheuristic algorithms is still lacking, despite the fact that prior research has focused on machine learning and deep learning related medical analysis issues associated with classification.

In order to enhance the precision of models, training performance, and mathematical time required for training for unbalanced collection of data, By correcting discrepancies in datasets and adjusting classifier parameter values, the study seeks to highlight the significant part that metaheuristic algorithms play in improving overall performance through understanding gleaned from experimental proof as well as complex data patterns.

1.1) Research contributions

The main contributions of this research are,

1) To create a hybrid Model by combining RNN and RBFN to increase diabetes prediction accuracy.

2) To improve data handling, using Dynamic Incremental Normalization and Evolutionary Gravitational Search for improved data preprocessing and feature selection.

3) To demonstrate the suggested model's effectiveness by outperforming traditional models on the PIMA Indian Diabetes Dataset.

2) Related Works

A machine learning model was developed by authors in [8] for the diabetes diagnosis procedure. The initial data for this approach came from the Pima Indian Diabetes dataset, which was obtained from the Kaggle Database Archive. Preprocessing was then used to eliminate any extraneous information. In this case, classification methods such as Support Vector Machine, Linear Regression and Naive Bayes Classifier were used. Improved rate of classification was attained by Support Vector Machine method. Various effective optimization techniques for various illness predictions were not included in this strategy, though. The investigators in [9] presented the decision tree model for diabetic illness prediction. Information Gain was used in this strategy to find relevant characteristics. Additionally, decision tree classifier was used to categorize Type 2 diabetes based on specific criteria. Although the proposed algorithm's classification performance was improved, the computational difficulty remains significant.

----

[8] K. De Silva, W.K. Lee, A. Forbes, R.T. Demmer, C. Barton, J. Enticott, Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis, Int. J. Med. Inform. 143 (2020) 104268, http://dx.doi.org/10.1016/j.ijmedinf.2020.104268.

[9] M. khan, B.K. Singh, N. Nirala, Expert diagnostic system for detection of hypertension and diabetes mellitus using discrete wavelet decomposition of photoplethysmogram signal and machine learning technique, Med. Nov. Technol. Devices 19 (2023) 100251, http://dx.doi.org/10.1016/J.MEDNTD.2023.100251.

The researchers in [10] presented a deep neural network classifier for the purpose of predicting diabetes. A variational autoencoder was used to obtain characteristics from the data that was entered, which was obtained from a central repository. Additionally, the softmax layer of the algorithm is used in the next phase of feature extraction to acquire the significant characteristics. Lastly, the constructed structure was fine-tuned using the Backpropagation technique. A multi-objective meta-heuristic approach for the diabetic diagnosis procedure was presented by authors in [11]. Here, the person with diabetes was identified using the K-means clustering algorithm. The multi-objective meta-heuristic technique was used in this framework to choose significant characteristics. To improve performance, this model did not, however, investigate the likelihood of diseases. A data mining method was developed in [12] for the classification of diabetes. The classification method in this model uses a Random Forest classifier. Despite failing to discover alternative categorization strategies for the acceptable results, this approach is more economical. The authors in [13] analyzed the PIMA data set to forecast diabetes. Out of all the machine learning methods used in this investigation, Naive Bayes classifier exhibited the highest accuracy.

Six machine learning methods were used in a different approach to predict diabetic conditions using the PIMA data set [14]. Random Forest was the most accurate of the algorithms considered for analysis. Additionally, a fuzzy rule-based model for classification was created in [15] for diabetes diagnosis.

----

[10] A. Prabha, J. Yadav, A. Rani, V. Singh, Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier, Comput. Biol. Med. 136 (2021) 104664, http://dx.doi.org/10.1016/j.compbiomed.2021.104664.

[11] M. Allwright, J.F. Karrasch, J.A. O’Brien, B. Guennewig, P.J. Austin, Machine learning analysis of the UK Biobank reveals prognostic and diagnostic immune biomarkers for polyneuropathy and neuropathic pain in diabetes, Diabetes Res. Clin. Pract. 201 (2023) 110725, http://dx.doi.org/10.1016/J.DIABRES.2023.110725.

[12] H.A. Abdulrahman, I.M. Olawale, S. Habeeb-Bello, S. Mohammed, A.J. Onumanyi, O.-O. Ajayi, Centers for disease control and prevention particle swarm optimization data set for diabetes classification, Open Sci. Framew. (2023) http://dx.doi.org/10.17605/OSF.IO/9JKQM.

[13] Y. Belsti, L. Moran, L. Du, A. Mousa, K.D. Silva, J. Enticott, H. Teede, Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model, Int. J. Med. Inform. 179 (2023) 105228, http://dx.doi.org/10.1016/J.IJMEDINF.2023.105228.

[14] S.S. Bhat, M. Banu, G.A. Ansari, V. Selvam, A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms, Healthc. Anal. 4 (2023) 100273, http://dx.doi.org/10.1016/J.HEALTH.2023.100273,

[15] C.B. Giorda, A. Rossi, F. Baccetti, R. Zilich, F. Romeo, N. Besmir, G. Di Cianni, G. Musacchio, Achieving good metabolic control without weight gain with the systematic use of GLP-1-RAs and SGLT-2 inhibitors in type 2 diabetes A machine-learning projection using data from clinical practice, Clin. Ther. 45 (2023) 754–761, http://dx.doi.org/10.1016/J.CLINTHERA.2023.06.006.

Using a comparable data set, authors in [16] deployed a deep neural network and achieved 98.05% accuracy. To exclude elements that can negatively impact the time required for execution, they employed feature extraction methods. Decision tree algorithm was used to determine feature relevance.

The work proposed in [17] chose the PIMA data set elements to incorporate into their forecasting algorithm using the Spearman correlation coefficient. They attempted to enhance the accuracy of diabetes prediction by employing this characteristic evaluation technique. Nevertheless, a number of features' compounding impact and unconventional association are poorly handled by the Spearman correlation algorithm. In an effort to create a strong framework, authors in [18] employed the Shapeley Augmented Approximation method to determine the significance of the characteristics used in the diabetes forecasting algorithm in the PIMA data set. For the small collection of data with only eight characteristics, their model's accuracy was 95.3%. This suggests that in order to choose the significant characteristics for this investigation, feature selection methods should also be advantageous.

The researchers in [19] employed five machine learning algorithms to forecast Type-II diabetes according to potential risk parameters that were discovered using multivariate logistic regression. However, this technique has certain drawbacks. For example, the authors neglected to verify for multi-collinearity before employing this model. Additionally, this approach makes the assumption that there is a direct correlation between the input parameters and the logarithmic chances of the resultant factors, but in reality, there may be various kinds of interactions. The work suggested in [20] integrated a machine learning method with network evaluation.

----

[16] A. Mujumdar, V. Vaidehi, Diabetes prediction using machine learning algorithms, Procedia Comput. Sci. 165 (2019) 292–299, http://dx.doi.org/10.1016/J.PROCS.2020.01.047.

[17] A. Nicolucci, L. Romeo, M. Bernardini, M. Vespasiani, M.C. Rossi, M. Petrelli, A. Ceriello, P. Di Bartolo, E. Frontoni, G. Vespasiani, Prediction of complications of type 2 Diabetes: A machine learning approach, Diabetes Res. Clin. Pract. 190 (2022) 110013, http://dx.doi.org/10.1016/J.DIABRES.2022.110013.

[18] S. Jangili, H. Vavilala, G.S.B. Boddeda, S.M. Upadhyayula, R. Adela, S.R. Mutheneni, Machine learning-driven early biomarker prediction for type 2 diabetes mellitus associated coronary artery diseases, Clin. Epidemiol. Glob. Health 24 (2023) 101433, http://dx.doi.org/10.1016/J.CEGH.2023.101433.

[19] S.M. Ganie, M.B. Malik, An ensemble machine learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators, Healthc. Anal. 2 (2022) 100092, http://dx.doi.org/10.1016/J.HEALTH.2022.100092.

[20] R. Cheheltani, N. King, S. Lee, B. North, D. Kovarik, C. Evans-Molina, N. Leavitt, S. Dutta, Predicting misdiagnosed adult-onset type 1 diabetes using machine learning, Diabetes Res. Clin. Pract. 191 (2022) 110029, http://dx.doi.org/10.1016/J.DIABRES.2022.110029.

Before using characteristics of networks and certain individual variables to predict type II diabetes, they first assigned patients to the disease system. Their best prediction accuracy was 92%, and adding network evaluation did not enhance machine learning effectiveness. A complete approach appropriate for large datasets is suggested in order to address the drawbacks of the state-of-the-art techniques for early type II diabetes diagnosis. Meta heuristic optimization is used for feature selection following data preprocessing. To cut down on computation time, only the most significant predictors are chosen from diabetic data sets, which contain a large number of features. Then, for effective diabetic prediction, a number of hybrid deep learning algorithms were used.

3) Proposed Methodology

The suggested methodology intends to improve diabetes prediction accuracy and efficiency by incorporating advanced data preprocessing, feature selection, and classification methods. Dynamic Incremental Normalization is used to standardize data distribution, reducing the influence of outliers while assuring effective learning. The Evolutionary Gravitational Search Algorithm is used to optimize the feature selection process. For classification, a hybrid model that combines Recurrent Neural Network (RNN) and Radial Basis Function Network (RBFN) is suggested. The RNN component detects sequential dependencies in the data, while the RBFN handles non-linear decision boundaries, yielding a strong predictive model. The architecture of proposed model is presented in Figure 1.

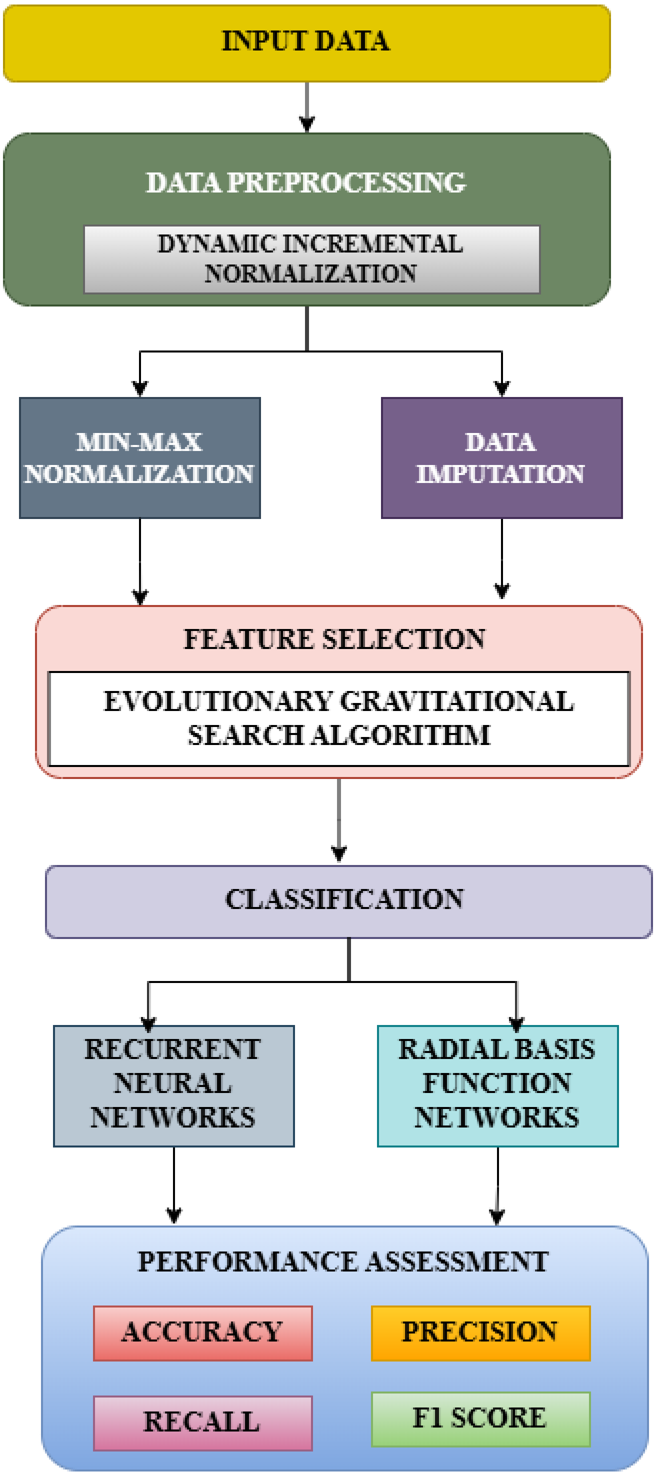

Figure 1 : Proposed Workflow

3.1) Data Preprocessing

Dynamic Incremental Normalization (DIN) is a sophisticated and unique type of normalization that was created specifically for data pre-processing. The DIN's potential stems from the fact that the log-linear algorithms implemented in this research are very appropriate for managing extensive and diverse datasets used in diabetes disease prediction. In order to create a more effective standardization operation, it is also responsive to the empirical characteristics of the source health information. Because the suggested DIN divides the data into multiple data portions and then applies Gaussian standardization procedures to each portion, it treats all of the information components in the supplied data set uniformly as compared to other available standardization techniques.

Here, the method makes use of the information's real distribution, ensuring that every portion is normalized with respect to the characteristics. This method uses a Gaussian normalization framework to build portions from the supplied healthcare data and determines appropriate boundaries. It successfully completes data resizing by transforming the information to obtain an average of 0 and a variance of 1. The DIN can detect changes in the data dispersion based on this statistical normalization process, which aids in the elimination of outliers and data biasing. The primary benefit of this approach is its adaptability. It can change the standardization and grouping components to provide a better, more ideal result. Furthermore, this method is highly helpful in healthcare settings where diagnostic criteria and population trends may differ among individuals' histories of illness.

Normalization and imputation are two preprocessing modules that make up the suggested DIN. First, every attribute in this framework is normalized to a range of zero to one using the minmax standardization approach as shown in equation (1),

$$m' = \frac{m - min(m)}{max(m) - min(m)}$$

(1)

Consequently, the acquired characteristics are converted into a distribution form using the Gaussian mixture model. as shown in equation (2),

$$m' = \frac{m - a}{v}$$

(2)

In the above equation, m denotes the current input sample, $a \wedge v$ denotes the average value and variance of the Gaussian component. Furthermore, incremental normalization is employed to segregate the feature into several subcomponents as shown in equation (3),

$$m'_{x,y} = \begin{cases} \partial_x \times (m_{x,y} - min(\mu_x))^{\sigma_x} if\ m_{x,y} \leq \alpha_x \\ \theta_x \times (m_{x,y} - min(\mu_x))^{\beta_x} if\ m_{x,y} > \alpha_x \end{cases} \quad (3)$$

In the above equation, $\partial_x, \theta_x, \sigma_x \wedge \beta_x$ are the factors used for resizing the portions, α_x denotes the maximum limit for segregated portion, μ_x is the actual segmented portion, $m_{x,y} \wedge m'_{x,y}$ represents the actual and normalized data samples. The solutions from all the segmented portions are combined using the formulation in equation (4),

$$m'_c = argmin_y(\vartheta(m'_{x,y})) \quad (4)$$

After that, the data imputation procedure is used to identify the values that were missing based on the characteristic the average or median estimates (δ) as shown in equation (5),

$$m_{imp} = \delta(m) \quad (5)$$

In order to identify the missing values using the incremental extrapolation model, a sequential imputation method is utilized to update the imputed data. The imputed values in standardized form can be represented as given in equation (6),

$$m_{imp}^{'} = \frac{m_{imp} - min(m)}{max(m) - min(m)}$$

(6)

Forecasts become more reliable and accurate when this preprocessing model is used since it significantly improves the standard of the incoming healthcare information. Because the DIN works by proactively dividing the information supplied into discrete parts based on its intrinsic dispersion characteristics. It uses the standard Gaussian normalization procedure for every portion, which is ideal for complex healthcare datasets because medical findings are greatly influenced by data volatility and complex patterns.

3.2) Feature selection using Evolutionary Gravitational Search algorithm

As an effective meta-heuristic, Evolutionary Gravitational Search Algorithm (EGSA) technique makes use of Newton's equations of gravity and velocity. In EGSA, search actors function as tangible components which communicate with one another; the size of these objects determines how well the outcome performs. Components may be drawn to one another by gravitational pull. The more delicate object is pushed up towards the object that is heavier by this pulling force. Heavy objects will travel more slowly, which is thought to be a superior option. This conduct guarantees the exploitation phase of EGSA. The vector F_k represents the location of each of the H search actors that are organized in the search environment with h dimensions as shown in equation (7),

$$F_k = \left(f_k^1, \ldots.. f_k^h\right) where k = 1, 2, \ldots., H \qquad (7)$$

Every vector value corresponds to a location of the actor in the particular direction, f_k^1 denotes the location of kth actor in the first dimension. The fitness of a search actor ($B_k(t)$) is computed using the equation (8) as,

$$B_k(t) = \frac{F_t - least(t)}{\sum_{l=1}^{H} (F_t - least(t))}$$

(8)

In the above equation, F_t denotes the fitness at t instant of time, $least(t)$ denotes the aggregate of the fitness of actors with least fitness. The search actors keep changing the locations in order to explore more regions in the search environment. The updated location $F_k(t+1)$ of a search actor is determined by adding the present location $F_k(t)$ and the consecutive velocity value $s_k(t+1)$ as given in equation (9),

$$F_k(t+1) = F_k(t) + s_k(t+1) \qquad (9)$$

To compute the consecutive velocity value, it is essential to add the acceleration $c(t)$ to the present value of velocity $s_k(t)$ as shown in equation (10),

$$s_k(t+1) = r \times s_k(t) + c(t) \qquad (10)$$

r is an arbitrary factor which holds a value either one or zero. The value of acceleration is computed based on the force and mass values of the components as represented in equation (11),

$$c_k(t)=\frac{\sum_{l=1}^{H}(p\times G_{kl}(t))}{B_k(t)}$$

(11)

In the above equation, p denotes an arbitrary value which is either one or zero. $G_{kl}(t)$ indicates the force that is acting upon component l on component k at time t. The value of force according to the gravitational law is calculated as per the formulation in equation (12) as,

$$G_{kl}(t)=\delta\frac{B_k(t)\times B_l(t)}{N^2}(F_l(t)-F_k(t)) \qquad (12)$$

In equation (12), gravitational constant is denoted as δ and N is the value of separation between component $k\wedge l$. To find the best characteristic groups, researchers have previously employed filtering techniques and adaptive algorithms. Information gain-based ranking of features is used to identify a characteristic group, which can be expressed by a vector of binary values. The aggregate number of characteristics chosen by the filtering process is the depth of the information gain-based characteristic assessment, where zero denotes the lack of an attribute and one denotes its existence. The equation (13) is used to update the location of the component and momentum at every cycle once the fitness function has assessed every characteristic value in light of the information gained.

$$u_{kl}^{t+1}=w\times u_{kl}^{t}+z_1\times rand\times\left(l\,pos_{best}-f_{kl}^{t}\right)+z_2\times rand\times\left(g\,pos_{best}-f_{kl}^{t}\right) \qquad (13)$$

In equation (13), u_{kl}^{t+1} and f_{kl}^{t}represents the value of velocity and position, w is a factor representing weights associated with the components, $l\,pos_{best}$ and $g\,pos_{best}$ denotes the optimal positions at local and global levels with rand, $z_1\wedge z_2$ corresponding to arbitrary values between 0 and 1. The best features are selected based on the velocity values as per the formulations in equations (14) to (16),

$$f_{kl}=\begin{cases}1, if\ rand<A(u_{kl})\\ 0, otherwise\end{cases} \qquad (14)$$

$$A(u_{kl})=\frac{1}{1+e^{-u_{kl}}} \qquad (15)$$

$$f_{kl}^{t+1}=f_{kl}^{t}+u_{kl}^{t+1} \qquad (16)$$

3.3) Classification using RNN and RBFN

To improve the accuracy of diabetes prediction, the hybrid model combines the characteristics of Radial Basis Function Networks (RBFN) and Recurrent Neural Networks (RNN). The idea behind this method is to take advantage of RBFN's skill at handling non-linear decision boundaries and RNN's capacity to describe temporal sequences in patient data. This combination is especially helpful for diabetes prediction since patient data frequently shows both non-linear patterns (like interactions between age, BMI, and glucose levels) and sequential dependencies (like time-series glucose readings).

3.3.1) Recurrent Neural Networks

RNNs are a type of artificial neural network that uses hidden states to retain a memory of prior inputs in order to identify patterns in sequential data. RNNs can remember information from earlier time steps because, in contrast to conventional feedforward networks, their connections loop back on themselves. Because of this, RNNs are especially well-suited for jobs involving time-series data, such forecasting the course of diabetes using past health information. The mathematical formulation of RNN model is represented in equation (17),

$$hd_k = \varphi(wt_{hd} hd_{k-1} + wt_a a_k + bs) \qquad (17)$$

In the above equation, hd, $wt, bs \wedge \varphi$denotes the hidden state, weights, bias and activation function. Sequential patterns in patient data, like changes in blood sugar levels over time, can be effectively captured by this model. It is appropriate for examining longitudinal health records since it preserves details about previous occurrences.

3.3.2 Radial Basis Function Network

Radial basis functions are used as activation functions in RBFNs, a type of artificial neural network. An input layer, a hidden layer with radial basis functions, and an output layer make up their three layers. RBFNs are effective for non-linear classification applications because the hidden layer converts the input data into a higher-dimensional space where it is linearly separable. The mathematical formulation of this model can be represented as shown in equation (18),

$$Ot_k(a) = \exp\left(-\frac{\text{¿} \vee a - d_k \vee \text{¿}^2}{2\theta^2}\right)$$

(18)

In the above equation, $Ot_k(a)$ denotes the output corresponding to the kth neuron in the hidden layer with d_k and θ denoting the center value and dispersion factor of the radial basis function. Complex interactions between risk factors, such as age, BMI, and glucose levels, can be handled by this approach. By facilitating a seamless transition between classes, it lessens overfitting.

3.3.3) Hybrid Classification model

RNN serves as a feature extractor in the hybrid approach, converting sequential patient data into a useful hidden state representation that records temporal dependencies. The RNN model is given sequential data. After processing this data in a sequential manner, the RNN creates a hidden state at the last time step that contains both recent and historical data. The input for the next RBFN step is this hidden state. The input for the next RBFN step is this hidden state. Gaussian radial basis functions are used to map the hidden state onto a higher-dimensional space. To generate the final forecast, RBFN calculates a weighted total of these modified inputs. For binary classification, the output layer is subjected to a sigmoid function.

4) Results and Discussion

4.1) Experimental setup

TensorFlow for the RNN component and Scikit-learn for the RBFN were used to set up the experimental setting for assessing the hybrid RNN-RBFN model for diabetes prediction using Python. To speed up training, the system configuration had an NVIDIA GPU with CUDA capability, an Intel Core i7 CPU, and 16 GB of RAM. In order to capture temporal relationships, the RNN was implemented with LSTM layers. The RBFN used Gaussian activation functions for classification, and its hidden states served as input features. The model was effectively trained using the binary cross-entropy loss function and Adam optimizer. Training time was greatly decreased by using GPU acceleration to speed up forward and backward propagation. Jupyter Notebook was used to manage the experimental environment for code execution and visualization. The accuracy, precision, recall, and F1 measures were used to assess the model's performance. This configuration guaranteed quick computation, excellent data processing, and accurate evaluation of the hybrid model's predictive power for diabetes.

4.2) Dataset Description

The dataset used in the experimentation is PIMA Indian Diabetes Dataset. For assessing deep learning models in diabetes prediction, this is a popular benchmark dataset. The National Institute of Diabetes and Digestive and Kidney Diseases gathered this dataset with the goal of forecasting the onset of diabetes in female patients with Pima Indian ancestry who are 21 years of age or older. Numerous physiological parameters associated with the risk of diabetes are described by the 8 clinical features and 768 samples, all of which are numerical. The binary target variable indicates if a patient has diabetes (1) or not (0). The features in the dataset are described in Table 1 and the corresponding correlation matrix is presented in Figure-2. The dataset can be accessed using the given link,
https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Table 1 Dataset details

Feature Name	Description	Measurement Unit
Pregnancies	The patient's total number of pregnancies.	Count in numbers
Glucose	Two hours following an oral glucose tolerance test, the plasma glucose concentration was tested.	mg/dL
Blood Pressure	Diastolic Blood Pressure	mm Hg
Skin Thickness	Triceps skinfold thickness (mm) as a measure of body fat.	mm
Insulin	2-Hour Serum Insulin	mu U/ml
BMI	Body Mass Index	Kg/ m²
Diabetes Pedigree Function	A function that calculates the risk of diabetes based on family history.	Count in numbers
Age	Age of the patient in years.	Count in numbers

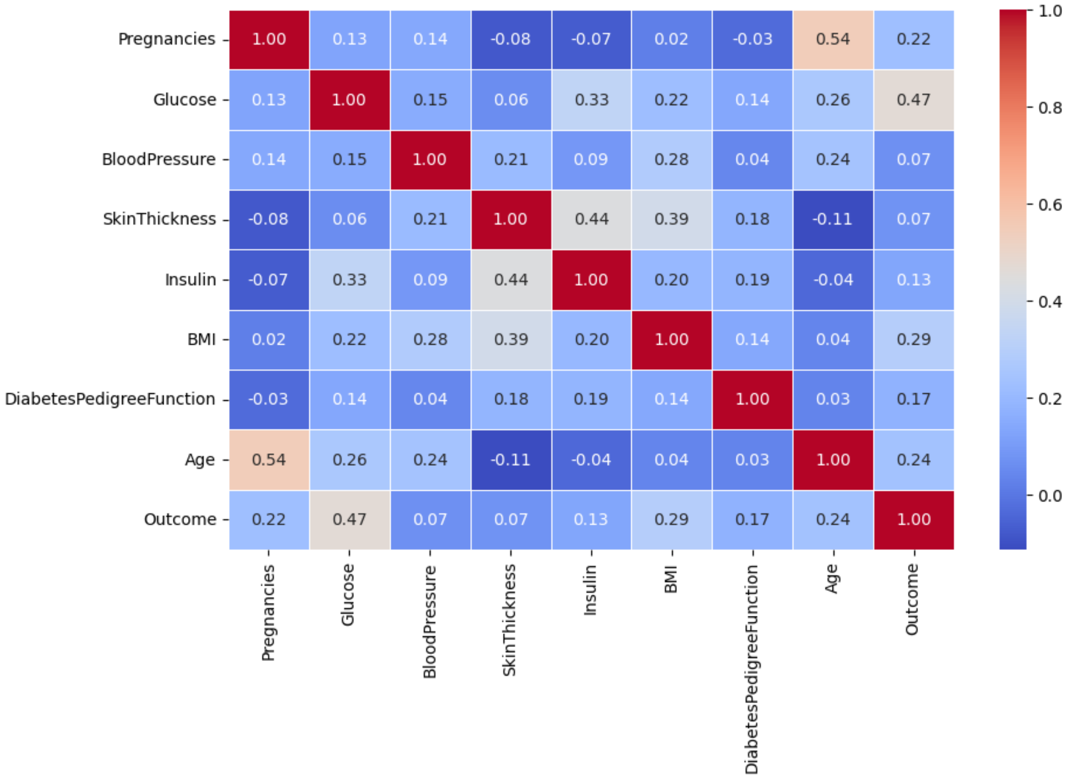

Figure 2: Correlation matrix for PIMA database

4.3) Performance assessment

Hyperparameter adjustment was crucial in improving the model's performance. The RNN component, with 64 hidden units and a learning rate of 0.001, effectively extracted temporal features, while the RBFN's 30 hidden neurons ensured that these features were properly mapped to the output layer. The RBFN's spread parameter (σ = 1.0) proved critical in balancing model complexity with generalization capabilities. Dropout regularization (rate: 0.3) helped to reduce overfitting, as indicated by the small difference in training and validation accuracy. To ensure the robustness and generalizability of the hybrid RNN-RBFN model, k-fold cross-validation was used throughout the experimental evaluation. In this work, a 10-fold cross-validation technique was used, with the PIMA Indian Diabetes Dataset partitioned into ten equal portions. In each iteration, 9-folds were utilized for training and 1-fold for testing, guaranteeing that every data point was used for both training and validation once. This strategy reduced the likelihood of overfitting while also providing a more trustworthy assessment of the model's performance on previously unknown data.

The average accuracy across 10 folds was 99.5% with a standard deviation of ±0.5%, demonstrating consistent performance and no notable variance between folds. This consistency demonstrates the model's ability to generalize effectively across different subsets of data. Furthermore, the precision and recall metrics varied just little, with average scores of 98.9% and 99.1%, indicating a balanced classification performance between diabetes and non-diabetic subjects. Cross-validation also offered information on the model's stability and the impact of specific attributes. For example, the findings revealed that parameters such as glucose and BMI consistently contributed to improved accuracy across multiple folds, reinforcing their value.

The hybrid model performed much better in each fold than standalone RNN and RBFN models, which had more variation and somewhat lower average accuracies (98.1% and 97.2%, respectively). Furthermore, cross-validation was useful in fine-tuning critical hyperparameters such as the number of hidden units in the RNN, the spread parameter in the RBFN, and the learning rate. Grid search and cross-validation confirmed that the chosen hyperparameters were optimal, which improved the model's prediction skills. The cross-validation findings highlight the hybrid approach's usefulness, demonstrating its dependability and robustness in predicting diabetes using the PIMA Indian Diabetes Dataset.

Table 1 presents the performance comparison of proposed model with conventional ML/DL models such as Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN), Multi-layer Perceptron (MLP), Recurrent Neural Network (RNN) and Radial Basis Function Network (RBFN).

Table 2 Performance Comparison with Conventional ML/DL models

Techniques	**Accuracy (%) ± Std**	**Precision (%) ± Std**	**Recall (%) ± Std**	**F1 Score (%) ± Std**
Support Vector Machine	94.7± 1.5	93.5± 1.7	94.2± 1.6	93.8± 1.5
Random Forest	96.2± 1.2	95.8± 1.3	95.9± 1.1	95.8± 1.2
Artificial Neural Network	97.5± 1.0	96.9± 1.2	97.2± 1.1	97.0± 1.0
Multi-layer Perceptron	97.8± 0.9	97.3± 1.0	97.5± 0.9	97.4± 0.9
Recurrent Neural Network	98.1± 0.8	97.6± 0.9	97.9± 0.8	97.7± 0.8
Radial Basis Function Network	97.2± 1.1	96.7± 1.2	96.9± 1.1	96.8± 1.1
Hybrid RNN-RBFN	99.5± 0.5	98.9± 0.6	99.1± 0.5	99.3± 0.5

The comparative examination of the models in the table demonstrates that the suggested Hybrid RNN-RBFN model outperforms traditional machine learning algorithms. The suggested model achieved 99.5% accuracy with a minimal standard deviation of ±0.5%, demonstrating great predictive performance and outstanding stability across multiple folds of cross-validation. The hybrid model achieved 98.9% ± 0.6% precision, outperforming existing models by efficiently limiting false positives. This is important for diabetes prediction to avoid unnecessary warnings. The recall score of 99.1% ± 0.5% indicates that the model accurately detects diabetes cases, lowering the chance of false negatives, which is crucial in medical diagnosis settings.

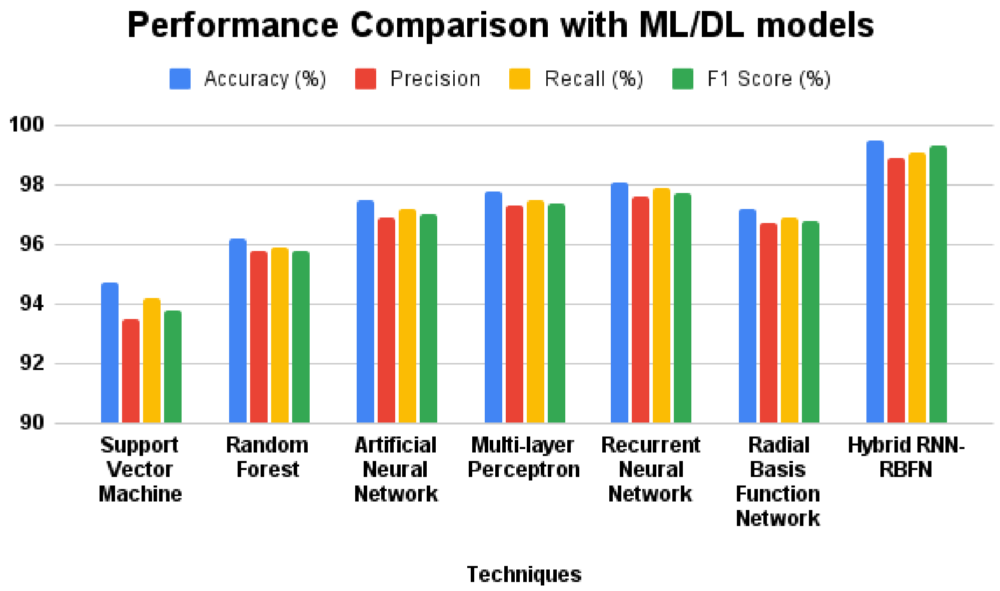

Figure 3: Performance Comparison with ML/DL models

The proposed model achieved the greatest F1-score (99.3% ± 0.5%), demonstrating its ability to balance sensitivity and specificity. The solo RNN and MLP models performed well, with accuracies of 98.1% ± 0.8% and 97.8% ± 0.9%, respectively. However, they fell short due to slightly higher variability and poorer precision and recall scores. The SVM and RBFN models had higher accuracies (94.7% ± 1.5% and 97.2% ± 1.1%, respectively), showing performance fluctuation due to hyperparameter sensitivity and difficulties in capturing complex data patterns. The Random Forest model achieved 96.2% ± 1.2% accuracy, but had lower performance than the hybrid model. This suggests that capturing both sequential dependencies (RNN) and non-linear decision boundaries (RBFN) in the proposed approach significantly improves predictive accuracy. Overall, the hybrid RNN-RBFN model's ability to successfully incorporate the capabilities of sequential learning and non-linear mapping has proven to be a solid solution for diabetes prediction, outperforming the other models with the least variability.

5) Conclusion

In this study, a unique hybrid technique integrating Recurrent Neural Network (RNN) and Radial Basis Function Network (RBFN) was established for accurate diabetes prediction. The suggested model used Dynamic Incremental Normalization for effective data preprocessing and the Evolutionary Gravitational Search Algorithm for optimal feature selection, resulting in greatly improved model performance and efficiency. Experimental results on the PIMA Indian Diabetes Dataset revealed the proposed method's superiority, with an impressive 99.5% accuracy, precision of 98.9%, recall of 99.1%, and F1-score of 99.3%. Comparative investigation indicated that the hybrid model beat classic classifiers such as SVM, Random Forest, ANN, MLP, standalone RNN, and RBFN in terms of accuracy and stability, demonstrating its capacity to handle complicated and nonlinear data patterns. However, the current research has certain drawbacks. The model's performance was assessed using a small and homogeneous dataset, which may not accurately reflect the diversity of real-world diabetes patients. Furthermore, the computational complexity of the hybrid model, which combines RNN and RBFN, may offer scaling issues for larger datasets and real-time applications. Future research could address these constraints by using more diverse and larger datasets, refining the model's design to reduce computing overhead, and investigating advanced optimization approaches. Despite these obstacles, the suggested strategy has tremendous potential for improving clinical decision-making in diabetes diagnosis.