A Hybrid RBF Neural Network and FCM Clustering for Diabetes Prediction Dataset

M. Khalil Gibran[1], Amir Saleh[2]
[1]Faculty of Computer Science & Information Technology, Universitas Muhammadiyah Sumatera Utara, Indonesia
[2]Department of Informatics Engineering, University of Prima Indonesia, Indonesia

ABSTRACT

This study aims to predict diabetes by combining the Radial Basis Function Neural Network (RBFNN) and Fuzzy C-Means (FCM) clustering methods. Diabetes prediction is an important part of research in an effort to prevent, manage, and reduce this type of disease. The FCM clustering method is used to group diabetes data into groups that have similar characteristics and obtain the final centroid. Then, the RBFNN method is used to build a predictive model using the center of each group as a reference point in the RBF function based on the centroid generated from the FCM clustering method. This step allows for modeling the non-linear relationship between health attributes and diabetes risk in more detail. In this study, the dataset obtained used input parameters regarding health data and risk factors for the disease. The goal of combining these methods is to develop a predictive model that can help identify individuals at high risk of developing diabetes. This hybrid approach has the potential to improve the effectiveness and accuracy of diabetes prediction. From the tests carried out, the proposed method obtained an accuracy of 92%, a precision of 90%, a recall of 92%, and an F1-score of 91%. By combining the clustering power of FCM clustering with RBF's ability to model non-linear relationships, this hybrid approach can make a good contribution to diabetes prediction and assist in efforts to prevent and control this disease.

Keyword : Diabetes Prediction, Radial Basis Function, Fuzzy C-Means, Hybrid

Corresponding Author:
Amir Saleh,
Department of Informatics Engineering
Universitas Prima Indonesia
Jl. Sampul, Sei Putih Bar., Kec. Medan Petisah, Kota Medan, Sumatera Utara
Email : amirsalehnst1990@gmail.com

Article history:
Received Jul 2, 2023
Revised Jul 29, 2023
Accepted Aug 30, 2023

1. INTRODUCTION

Diabetes mellitus is a very dangerous disease because it contributes to other deadly diseases such as kidney, heart, and nerve damage (Butt et al., 2021)(Lubis et al., 2019). Predictive methods are an important aspect of efforts to prevent, early diagnose, and manage diabetes. In recent years, machine learning and data mining methods have been widely used and proven effective in developing diabetes prediction models. In addition, other techniques for predicting diabetes have been used with satisfactory results, one of which is the artificial neural network (ANN). Through artificial neural network techniques, we can design and implement complex medical processes in software. The software system will become more effective and efficient in various health fields, including predicting, treating, diagnosing, and assisting health teams and the general public (Pradhan et al., 2020)(Al-Khowarizmi et al., 2017).

One of the neural network methods that can be used to predict diabetes is the RBFNN. However, there are limitations to using this method, which are dependent on the selection of the right parameters. Parameters such as the centroid, width, and initial weight of the RBFNN must be carefully adjusted to obtain optimal results (Saleh et al., 2019). If these parameters are not chosen properly, the performance of the RBFNN can suffer significantly. The solution can be solved by combining methods, such a fixing the initial weight, which gives better performance. The combination of the RBFNN method with BAT Optimization obtains better performance compared to conventional RBFNN networks in terms of accuracy, specificity, sensitivity, complexity, and time (Cheruku et al., 2017).

Another technique that can be used is predictive analysis, which involves combining various machine learning algorithms, data mining techniques, and statistical methods to obtain higher predictive performance. This technique is widely applied to study past data to gain knowledge and predict future events. By applying predictive analysis techniques to health data, significant decisions can be made and predictive outcomes improved (Mujumdar & Vaidehi, 2019).

In this study, we propose an approach by combining the FCM clustering method with the RBFNN for the prediction of diabetes. This approach takes advantage of both methods to increase the effectiveness and accuracy of diabetes prediction. The FCM method is a clustering technique that is able to overcome data complexity and uncertainty in traditional clustering algorithms. In addition, FCM allows for the degree of data membership in each cluster, so that data that is ambiguous or not completely homogeneous can be put into the most appropriate group (Hossein-Abad et al., 2020). On the other hand, the RBFNN is a powerful neural network method for modeling non-linear patterns and relationships in data (Chai et al., 2019). The RBFNN method is able to handle data complexity by changing the data representation into a more abstract feature space using non-linear basis functions.

The initial stage of this research was carried out using the clustering method using FCM to classify diabetes data into homogeneous groups. Then, the final result of this method is a data center that can be used as an initial weight initialization in the RBFNN algorithm. In the next stage, the RBFNN method will be trained to model patterns and relationships between data attributes in each group. The dataset used in this study contains medical information and clinical data relevant to diabetes obtained from the Kaggle dataset. In the final stage, we evaluate and compare the prediction results obtained with the traditional RBFNN method to see the improvement in the results obtained. The main objective of this study is to predict the risk of diabetes in patients by identifying patterns and relationships between these variables. It is hoped that this research will make a good contribution to the development of a more accurate and efficient diabetes prediction model.

2. RESEARCH METHOD

The methodology in this study provides general guidance on the steps to be taken in research to implement the RBFNN-FCM hybrid approach for the prediction of diabetes. The steps taken can be seen in Figure 1 below.

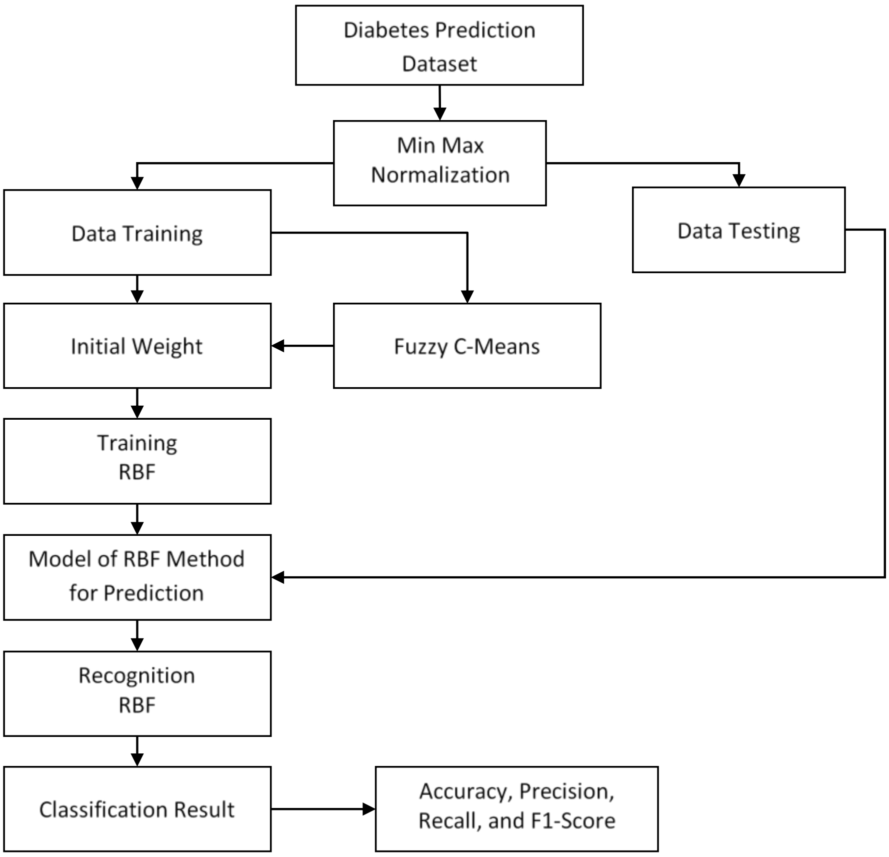

Figure 1. The proposed method in the diabetes prediction dataset

A. Data Collection

The dataset used in this study was obtained from the Kaggle Dataset, which is the Diabetes prediction dataset. The data is obtained based on a collection of medical and demographic data from patients and their diabetes status (positive or negative). There are attributes such as gender, age, hypertension, body mass index (BMI), heart disease, smoking history, blood glucose levels, and HbA1c levels. The dataset obtained is used to build a machine learning model to predict patient diabetes based on medical history and demographic information. Information about the dataset can be seen in Table 1 below.

Table 1. Information the dataset

No.	Attributes	Value
1.	Gender	• Famale • Male
2.	Age	0.08 – 80
3.	Hypertension	• 0 • 1
4.	Heart Disease	• 0 • 1
5.	Smoking History	• No Info • Never • Former • Current • Not Current
6.	BMI	10 – 95.7
7.	HbA1c	3.5 – 9
8.	Blood Glucose	80 – 300
9.	Diabetes	• 0 • 1

B. Data Pre-Processing

The initial stage of this study was to clean the data to remove invalid entries, fill in missing values, and normalize attributes to obtain comparable values. The normalization technique used is the min-max method, which can be calculated using the following equation 1 below (Patro & sahu, 2015).

$$A' = \left(\frac{A - \min value\ of\ A}{\max value\ of\ A - \min value\ of\ A}\right) * (D - C) + C \tag{1}$$

Where,

A’: Min-Max Normalized data
C, D: Predefined boundary
A: The value of original data

C. Fuzzy C-Means (FCM) Clustering

The FCM algorithm in this study is used for grouping data and obtaining the optimal centroid. The use of this method requires parameters, such as the desired number of clusters and the initial degree of membership. Then, update the cluster centroid and degree of membership based on the resulting FCM objective function. This process will be carried out until convergence is reached according to the value given. The end result of this method is to obtain the final centroid that will be used in the RBF as the initial weight of the network, which is the contribution offered in this research. The equation used to find the final centroid value in the FCM algorithm can be seen in equation 2 below (Dini & Maarif, 2022).

$$V_{kj} = \frac{\sum_{i=1}^{n} (\mu_{ki})^{w} X_{ij}}{\sum_{i=1}^{n} (\mu_{ki})^{w}} \tag{2}$$

Where,

V_{jk}: cluster center on the k-th and j-th
μ_{ki}: u matrix (partition data) in the k-th and i-th
X_{ij}: data on the j-th and i-th
w: weighting

D. Radial Basis Function (RBFNN) Neural Network

The RBFNN method implements mathematical functions that are often used in modeling and data analysis. The function will be determined by the radial distance between the input point and the predefined data center. The RBF can generate output values based on the distance between the input point and the center, taking into account a certain radial function. The RBF network has strong functional approximation capabilities and is proven to be an effective tool for modeling nonlinear processes (Yang et al., 2022). In general, RBFNN consists of three layers: the input layer, the hidden layer, and the output layer. The mathematical equation of the RBFNN output can be calculated using equation 3 below (Yang et al., 2022).

$$\hat{y}(t) = \sum_{j=1}^{m} w_j(t).\,\theta_j(t) \qquad (3)$$

Where,

m: number of hidden neurons
$w_j(t)$: weight between the j-th hidden layer and the output layer
$\theta^j(t)$: output from the j-th hidden layer neuron

To find the value of the hidden layer neurons, you can use the following equations: 4 and 5 below.

$$\theta_j(t) = e^{-\frac{\left\|x(t)-c_j(t)\right\|^2}{2\sigma_j^2(j)}} \qquad (4)$$

Where,

$$c_j(t) = \left[c_{1,j}(t), c_{2,j}(t), \dots, c_{n,j}(t)\right]^T \qquad (5)$$

Where,

x(t): input vector
$c_j(t)$: centroid of hidden j-th neurons
n: dimension of the input vector
$||x(t)-c_j(t)||$: Euclidean distance between x(t) and $c_j(t)$
$\sigma_j(t)$: width of the jth hidden neuron

In general, in the RBFNN, the initial weight is determined by generating random values. This value greatly affects the results of the RBFNN and reduces the performance obtained when an incorrect determination is made (Alzaeemi et al., 2019). One solution is to combine other methods in the process of determining the value. In this study, this value will be determined by the FCM algorithm using the proposed hybrid technique. The FCM method will find the optimal center value based on the iterations that are run, and the final result can be used as the initial weight on the RBF network.

E. Model Evaluation

Evaluation of algorithm performance on diabetes prediction involves using evaluation metrics that are appropriate for a binary classification problem. Several evaluation metrics commonly used to measure model performance in predicting diabetes are Accuracy, Precision, Recall, and F1-Score. The equation for finding these values can be found using the following equations: 6, 7, 8, and 9 below (Ramli et al., 2022)(Tasnim et al., 2022).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (6)$$

$$Precision = \frac{TP}{TP+FP} \qquad (7)$$

$$Recall = \frac{TP}{TP+FN} \qquad (8)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision \times Recall} \qquad (9)$$

Where,

TP: True Positive
TN: True Negative
FP: False Positive
FN: False Negative

The results of the hybrid method obtained will be compared with the prediction results of the traditional RBFNN method to determine the improvement of the proposed model. This hybrid approach takes advantage of the power of Fuzzy C-Means Clustering in data grouping and segmentation as well as the ability of RBFNN to model complex non-linear relationships between attributes and diabetes risk. With this combination, it is hoped that diabetes prediction will be more accurate and effective in identifying individuals at high risk of developing diabetes based on their characteristics and health attributes.

3. RESULTS AND DISCUSSION

In this study, a hybrid method will be implemented by combining the RBFNN and FCM approaches to complete the detection of diabetes. The RBFNN method is used as a model to classify data, while the FCM method is used to produce initial weights to improve classification performance. In addition, the test will be compared with the usual RBFNN method to find out the improvement from the proposed approach. The values of the learning parameters in the hybrid and regular RBFNN methods are the same, and the difference is only made in determining the initial weight of the RBFNN. In the usual RBFNN, the way to do this is to enter as many random values as the input matrix and data class allow. As for the proposed method, this process is carried out using the FCM method. The FCM parameters used in this study can be described as follows:

- c = 2
- threshold = 10^{-5}
- w = 2
- $P_0 = 0$

Meanwhile, the RBFNN parameters used in this study can be described as follows:

- k = 10
- lr = 0.1
- epochs = 100
- s = 1.0
- m = 2

After determining the parameter values needed for the prediction process, the next step is to carry out the training process by involving as many as 3,500 patients, whether they have diabetes or not. This is done to obtain a network model, which will then be tested on a dataset of 1,500 to obtain predictive results. From testing the two methods used, the results obtained can be seen in Table 2 below.

Table 2. Methods in diabetes prediction

Methods	Accuracy	Presicion	Recall	F1-Score
RBFNN	89%	88%	89%	89%
RBFNN-FCM	92%	90%	92%	91%

The RBFNN and FCM hybrid method produces better diabetes prediction accuracy, and 1500 patients have been tested for diabetes prediction. From the prediction results, 1380 patients were classified correctly, while 120 patients were classified incorrectly. Based on the tests performed, the results obtained can be described as follows:

- Accuracy: The hybrid method of RBFNN and FCM yields an accuracy of 92%, which means that 92% of the total test data is classified correctly. Whereas using the usual RBFNN, the accuracy obtained is 89%.
- Precision: The precision of this hybrid method reaches 90%, meaning that 90% of the positive predictions made are actually cases of diabetes. Whereas using the usual RBFNN, the precision obtained is 89.
- Recall: Recall of this hybrid method reaches 92%, meaning 92% sensitivity in cases of properly detected diabetes. Meanwhile, using the usual RBFNN, the recall obtained was 89%.
- F1-Score: The RBFNN and FCM hybrid methods achieve an F1-Score value of 91%, while using the usual RBFNN, the F1-Score value obtained is 89%.

The RBFNN and FCM hybrid methods combine two different approaches. The RBFNN method is used to build a classification model using a radial function that can map data into feature space. Meanwhile, FCM is used to produce cluster centers that represent centers of possible data membership in a particular group. By combining these two methods, the hybrid method tries to take advantage of the advantages of each method to improve prediction performance. Based on the combination of the two methods, the improvement obtained can be seen in Figure 2 below:

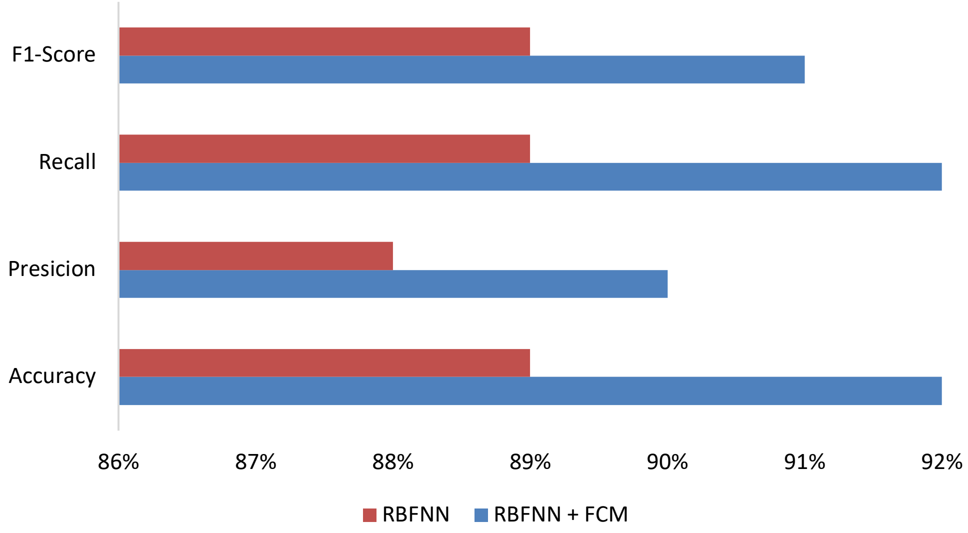

Figure 2. Results of a comparison of methods in diabetes prediction

On Figure 2, it can be seen that the increase in accuracy that occurred using the proposed method was 3%, precision increased by 2%, recall increased by 3%, and F1-Score increased by 2%. The improvement obtained using the proposed method occurs in each evaluation method used. This states that the method is able to improve the performance of the RBFNN method in predicting diabetes.

The RBFNN method in this study has the ability to produce optimal performance from data classification. The addition of FCM helps in differentiating relevant characteristics between patients who have diabetes and those who do not have diabetes at an early stage. With good FCM ability, the model can have a better representation of determining the initial weight and improve its predictive ability. RBFNN and FCM hybrid methods can provide better class separation between diabetic and non-diabetic patients. By considering the information from FCM about the membership of the data in certain groups, the model can obtain a better representation of the relationship between the features and the target class. This helps improve predictive accuracy by better separating diabetic patients from non-diabetics.

Through the combination of RBF and FCM, this hybrid method can provide more reliable and valid prediction results. The FCM approach using fuzzy membership helps to account for the degree of uncertainty in classification, while the RBF provides a robust framework for modeling the relationship between features and target classes. By combining the two, the prediction results become more solid and are able to overcome ambiguity in the data. At the development stage, it is important to carry out further evaluation and optimization of this hybrid method. In addition, parameter selection and model optimization also need to be done to obtain better results. This process can involve experimenting with various parameter configurations and other techniques.

4. CONCLUSION

In this study, a hybrid method of RBFNN and FCM clustering was used to predict diabetes in 1500 patients tested. The prediction results show that the hybrid method produces better accuracy than the usual RBFNN. The RBFNN and FCM hybrid methods produce an accuracy of 92%, a precision of 90%, a recall of 92%, and an F1-Score of 91%. Whereas using the usual RBFNN, the accuracy obtained is 89%, precision is 89%, recall is 89%, and the F1-Score is 89%. Through the combination of RBF and FCM, this hybrid method can provide more reliable and valid prediction results. The FCM approach using fuzzy membership helps to account for the degree of uncertainty in classification, while the RBF provides a robust framework for modeling the relationship between features and target classes.

REFERENCES

Al-Khowarizmi, A., Sitompul, O. S., Suherman, S., & Nababan, E. B. (2017). Measuring the Accuracy of Simple Evolving Connectionist System with Varying Distance Formulas. *Journal of Physics: Conference Series*, *930*(1). https://doi.org/10.1088/1742-6596/930/1/012004

Alzaeemi, S., Mansor, M. A., Mohd Kasihmuddin, M. S., Sathasivam, S., & Mamat, M. (2019). Radial basis function neural network for 2 satisfiability programming. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(1), 459–469. https://doi.org/10.11591/ijeecs.v18.i1.pp459-469

Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications. *Journal of Healthcare Engineering*, *2021*. https://doi.org/10.1155/2021/9930985

Chai, S. S., Wong, W. K., Goh, K. L., Wang, H. H., & Wang, Y. C. (2019). Radial basis function (RBF) neural network: Effect of hidden neuron number, training data size, and input variables on rainfall intensity forecasting. *International Journal on Advanced Science, Engineering and Information Technology*, *9*(6), 1921–1926. https://doi.org/10.18517/ijaseit.9.6.10239

Cheruku, R., Edla, D. R., & Kuppili, V. (2017). Diabetes classification using Radial Basis Function Network by combining cluster validity index and BAT optimization with novel fitness function. *International Journal of Computational Intelligence Systems*, *10*(1), 247–265. https://doi.org/10.2991/ijcis.2017.10.1.17

Dini, D. F. R., & Maarif, S. (2022). Desimal : Jurnal Matematika. *Desimal: Jurnal Matematika*, *V*(1), 31–102. https://doi.org/10.24042/djm

Hossein-Abad, H. M., Shabanian, M., & Kazerouni, I. A. (2020). Fuzzy c-means clustering method with the fuzzy distance definition applied on symmetric triangular fuzzy numbers. *Journal of Intelligent and Fuzzy Systems*, *38*(3), 2907–2950. https://doi.org/10.3233/JIFS-180971

Lubis, A. R., Prayudani, S., Lubis, M., & Al-Khowarizmi. (2019). Analysis of the Markov Chain Approach to Detect Blood Sugar Level. *Journal of Physics: Conference Series*, *1361*(1). https://doi.org/10.1088/1742-6596/1361/1/012052

Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 292–299. https://doi.org/10.1016/j.procs.2020.01.047

Patro, S. G. K., & sahu, K. K. (2015). Normalization: A Preprocessing Stage. *Iarjset*, 20–22. https://doi.org/10.17148/iarjset.2015.2305

Pradhan, N., Rani, G., Dhaka, V. S., & Poonia, R. C. (2020). Diabetes prediction using artificial neural network. *Deep Learning Techniques for Biomedical and Health Informatics*, *121*, 327–339. https://doi.org/10.1016/B978-0-12-819061-6.00014-8

Ramli, N. E., Yahya, Z. R., & Said, N. A. (2022). Confusion Matrix as Performance Measure for Corner Detectors. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *29*(1), 256–265. https://doi.org/10.37934/araset.29.1.256265

Saleh, A., Tulus, T., & Efendi, S. (2019). *Analysis of Accurate Learning in Radial Basis Function Neural Network Using Cosine Similarity on Leaf Recognition*. https://doi.org/10.4108/eai.20-1-2018.2281924

Tasnim, A., Saiduzzaman, M., Rahman, M. A., Akhter, J., & Rahaman, A. S. M. M. (2022). Performance Evaluation of Multiple Classifiers for Predicting Fake News. *Journal of Computer and Communications*, *10*(09), 1–21. https://doi.org/10.4236/jcc.2022.109001

Yang, Y., Wang, P., & Gao, X. (2022). A Novel Radial Basis Function Neural Network with High Generalization Performance for Nonlinear Process Modelling. *Processes*, *10*(1), 1–16. https://doi.org/10.3390/pr10010140