Diabetes Classification using an Expert Neuro-fuzzy Feature Extraction Model

P. Bharath Kumar Chowdary[1]*
Research Scholar, Department of Computer Science and Engineering, BIST, Bharath Institute of Higher Education and Research (BIHER), India

Dr. R. Udaya Kumar[2]
Research Supervisor, Professor, Department of Information Technology, BIST, Bharath Institute of Higher Education and Research (BIHER) Institution, India

***Abstract*—Diabetes is one of the challenging diseases prevailing in recent times. Due to the incompleteness, uncertainty and imprecise details, classification of diabetes using machine learning algorithms is turning out to be even more challenging. The efficiency of the classification model is influenced by the data present in the dataset. This study enhances the classification of diabetes by using a Neuro-Fuzzy model with special attention to Feature Extraction. The main goal of the present study is to enhance the diabetes prediction technique that helps the medical practitioners to easily identify the disease and diagnose it appropriately to reduce several complications that diabetes may cause to the patient in the future. The proposed model initially applies fuzzification on diabetes data to produce membership values. Later the membership values are examined by the proposed model to check the contribution of the features in diabetes classification. The feature extraction algorithm passes the significant features to a neural network after the features are extracted. The proposed model is tested on standard PIMA diabetic dataset to evaluate the performance. The proposed model is able to outperform all the existing machine learning algorithms.**

Keywords—Diabetes; neuro-fuzzy model; feature extraction; artificial neural network

I. Introduction

Diabetes is now one of the most affecting diseases on the human race. Diabetes occurs mainly due to the insufficient production of insulin in the body. As per the details of the World Health Organization (WHO), diabetes is rapidly increasing in developing and underdeveloped countries. It is expected that by 2030, the severity of the disease increases very drastically, becoming a dreadful disease that will lead to the death of many human beings [1], [2]. Diabetes comes in three variations; namely Type I, Type II and Gestational.

Type I diabetes occurs mostly in children and is very rare, Gestational diabetes occurs during pregnancy, but both of these are very rare. Type II diabetes is more common in humans [3]. Type II diabetes causes a serious effect on the health of an individual and at present, there is a lot of research going to predict diabetes at the early stages using various models [4], [5]. Type II diabetes models require appropriate algorithms to efficiently detect the disease and thus help physicians to diagnose the disease early [6]. Early diagnosis [7] of the disease helps to overcome the impact that diabetes causes on various organs in the human body like kidneys, heart, eyes, etc. The disease prediction models that handle the diabetes data often face issues like noisy, missing, irrelevant and inconsistent data [15], [16].

The performance of the model depends on the quality of the diabetes data presented to the model and hence the researcher must supplement accurate data to the classifier for effective disease prediction [17]. In the machine learning domain, classification is an important task as it derives knowledge to handle real-world applications [18], [19].

Classification constructs a model to predict the target class of the data accurately. Classification models like artificial neural networks (ANN) do not work efficiently to produce high accuracy due to their slow convergence rate, also suffer from the local minima problem and is a very high computational model. Although ANN is adaptive, it requires a high amount of time to produce the result because of its complex structure. ANN is not ideal to deal unclear, imprecise details [7].

To address the various issues of ANN, this paper aims to develop an ANN-based fuzzy model (ANNFM) that converts the numeric features into appropriate linguistic terms such as very low (VL), low (L), medium (M), high (H), and very high (VH). Each feature in the dataset is converted into a suitable membership value based on the five linguistic terms mentioned above. Thus all the features in the dataset are converted into linguistic features. Fuzzy logic enhances the ANN model to deal with the uncertainty problem by giving membership values to all the features. The proposed model performs classification by using fuzzy logic and the ANN model. The majority of classification models reduce the features at the pre-processing stage. This leads to loss of information and prediction capability will be affected if the features are reduced in the training process. Hence the proposed approach considers all the features initially and applies fuzzification on the features to determine the significant features for efficient decision making. The main motive of this research is to develop a model that integrates the linguistic terms of fuzzy logic and ANN and to use an efficient feature extraction algorithm to classify the data.

The rest of the paper is organized as follows: the second section covers the existing work in this domain. Section 3 presents the proposed approach for diabetes classification based on fuzzy linguistic terms and artificial neural networks with special attention to feature extraction. Section 4 presents the results of the proposed approach and the state-of-art

*Corresponding Author

diabetes classification approaches. Section 5 concludes the paper with a direction towards possible future aspects.

II. Related Work

This section covers the related work in the domain of diabetes classification and management. This section also explains how the existing models predicted different alignments using neuro-fuzzy models.

This work [7] presents an adaptation of neuro-fuzzy inferences to perform the classification of electrocardiogram signals. This model integrated the benefits of fuzzy inference system with the neural network for better classification of electrocardiogram signals. This work [8] presents a machine learning paradigm to classify diabetic data. The work focused on applying classification techniques like linear discriminant analysis, Naive Bayes to classify the PIMA dataset. The authors did not focus on efficient preprocessing techniques. The performance of various classifiers using various evaluation metrics is reported in this paper. But due to inefficient handling of missing data the results are not higher on the PIMA dataset.

The authors in [9] performed the classification of diabetes using the Levenberg-Marquardt learning algorithm. The authors constructed a neural network to perform the classification of the PIMA diabetes dataset. The algorithm is dynamically applied to the neural network to calculate the sensitivity and specificity of the developed model. The authors did not handle the missing values of the dataset properly and hence the results of this model are not promising.

The work [10] presents a model to classify diabetic data. This model aims to integrate Ant Colony-based optimization model with fuzzy rules to diagnosis the diabetic data. The optimization proved effective in classifying the data and achieved better results. This work [11] presented an ensemble system to diagnose diabetes using J48 and Adaboost techniques. The main work this paper focused on is to use the rules to help undiagnosed individuals to reduce the risk of diabetes incidence. The main aim of this work is to classify different adult groups of Canada and help the physicians to carry out the research. The works [12], [14] are also a machine learning approach to predict diabetes on the PIMA dataset.

The authors [13] proposed a neural network-based model using radial basis function to classify the diabetes data. The model comprises hidden layers also to increase the performance of the classification. The model uses the Bat-based clustering approach to find the number of neurons required in the hidden layer. The works [19], [20] focused on diagnosing kidney disease in the patients who are affected with the diabetes using classification. The authors [21] worked on the adverse effects that diabetes disease cause with respect to cardiovascular and kidney diseases.

Few models based on deep learning [15], [16] also stressed the importance of analyzing the diabetes at the early stages. All the discussed models in this section focused on diabetes classification only. The machine learning algorithms also paid attention to classify diabetes data. The discussed models paid little attention to the data pre-processing and fuzzification aspects. Hence in the next section, an efficient model to perform data preprocessing efficiently and to apply the fuzzy linguistic parameters to enhance the artificial neural networks for classifying type-II diabetes on the PIMA dataset is presented.

III. Materials and Methods

This section provides the detailed description of the proposed work and the dataset used.

A. Assigning Fuzzy Membership Values to the Features in the Dataset using Fuzzification and Performing Classification of Data

The initial phase is to perform the fuzzification process on the data. Each entry in the dataset consists of many features. In the first phase, this work converts the feature into a linguistic term like VL, L, M, H and VH.

If dataset is having ‘P’ records with each ‘n’ features per record,

$$P_i = [\, f_{i1},\, f_{i2}, ...f_{in}\,] \tag{1}$$

Where ‘i’ represents the ith record in the dataset and f_{ij} represents the jth feature of the record.

For each feature in the record membership values are assigned using the π-type function. The π-type assigns fuzzified values based on the five linguistic terms. Hence, the feature vector contains 5*n fuzzified features, if there are ‘n’ features per record.

$$f_{ij} = [\mu_{VL}(f_{ij}), \mu_{L}(f_{ij}), \mu_{M}(f_{ij}), \mu_{H}(f_{ij}), \mu_{VH}(f_{ij})] \tag{2}$$

The data is transformed into fuzzy membership values initially. Once each feature is assigned with five membership values, the size of the feature vector grows significantly which increases complexity. To overcome this problem, the present work uses principal component analysis to store all the significant features and discard the unnecessary features. Moreover, the data in the dataset has many missing values. This paper uses imputation techniques to fill the missing data and after the fuzzification process, all the significant features are restored. The procedure of the proposed methodology is shown in Fig. 1. There are two phases in the proposed approach. Fig. 1 shows the detailed step by step working principle of the paper.

The first phase performs two tasks, a) fuzzification of data and b) feature extraction. The second phase handles the neural network aspects of the model. After the feature extraction model, a suitable neural network is built to perform classification of data. The process of classification is shown in Fig. 2. The input to the ANN is membership values of the feature vector. The initial weights of the network are in the range (0, 1). The input layer has as many nodes equivalent to the number of features after feature reduction. The output layer has nodes equivalent to the number of the classes in the dataset.

B. PIMA Dataset

This dataset has records of 768 patients and for each record there are 8 features and a class label, specifying whether the patient is diabetic or non-diabetic. The description of various attributes of the dataset is mentioned in Table I.

The dataset has many missing values. The proposed model has used imputation techniques to fill the missing values in the dataset. For example, the Glucose column in the dataset has 374 and the skin thickness column has 227 missing values. Along with these two columns, there are some other missing values in the dataset also. This work applies the imputation technique to fill the missing values. Later the features in the dataset are assigned fuzzy membership values and after this, the significant features are extracted using principal component analysis. The feature vectors of all the records of the PIMA dataset are obtained at this point.

The missing values in the dataset are shown in Fig. 3. From the figure, it is observed that the skin thickness and insulin are having more missing values.

In this paper imputation techniques are used to fill the missing values of the dataset for efficient classification. After the imputation, the correlation of the attributes in the dataset is shown in Fig. 4.

From Fig. 4, it is observed that the attributes pregnancies and age are more correlated. BMI and skin thickness, Glucose levels and Insulin are also more correlated.

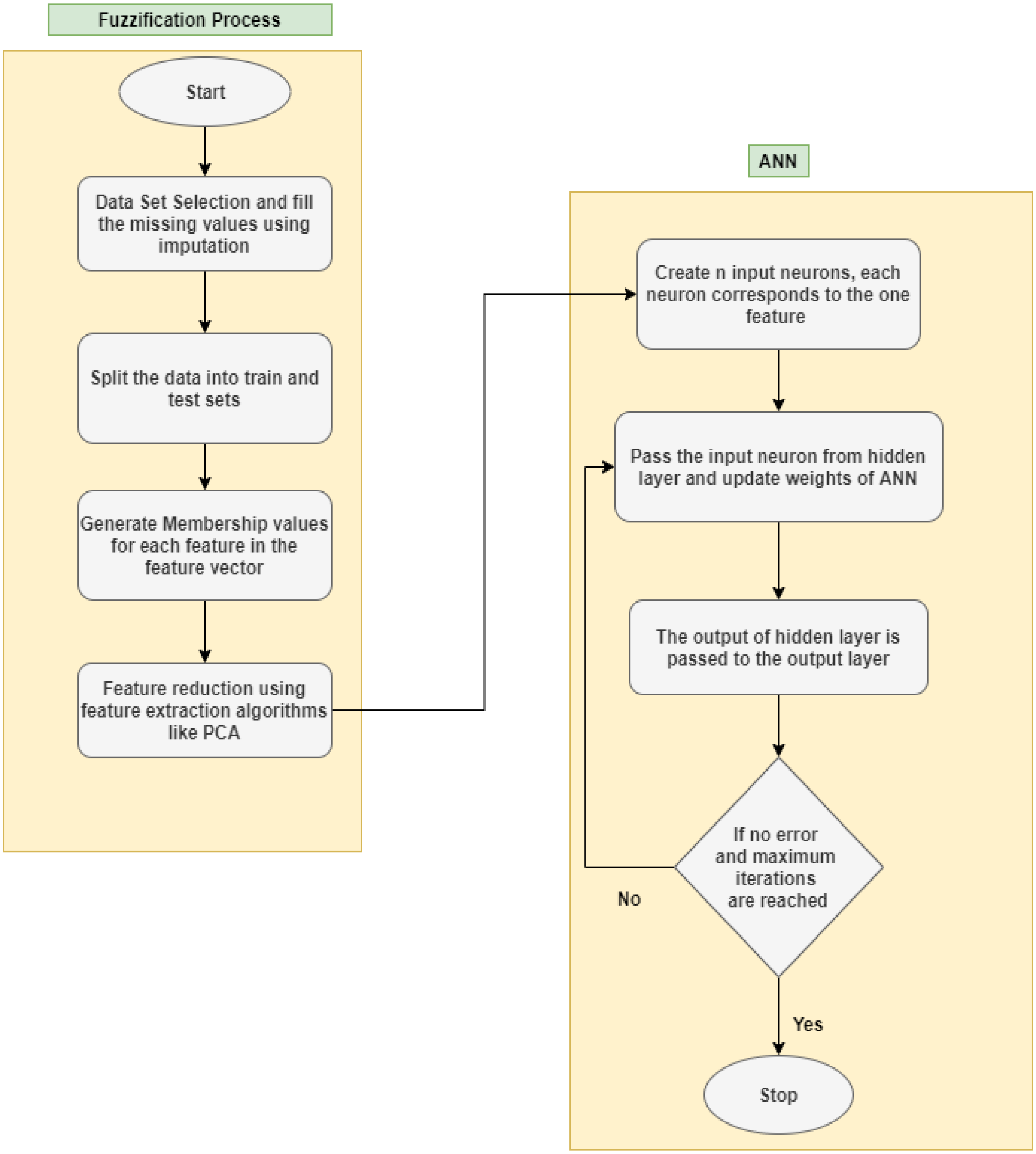
Fuzzification Process
Start
Data Set Selection and fill the missing values using imputation
Split the data into train and test sets
Generate Membership values for each feature in the feature vector
Feature reduction using feature extraction algorithms like PCA
ANN
Create n input neurons, each neuron corresponds to the one feature
Pass the input neuron from hidden layer and update weights of ANN
The output of hidden layer is passed to the output layer
If no error and maximum iterations are reached
No
Yes
Stop

Fig. 1. Architecture of Proposed Model to Classify Diabetes.

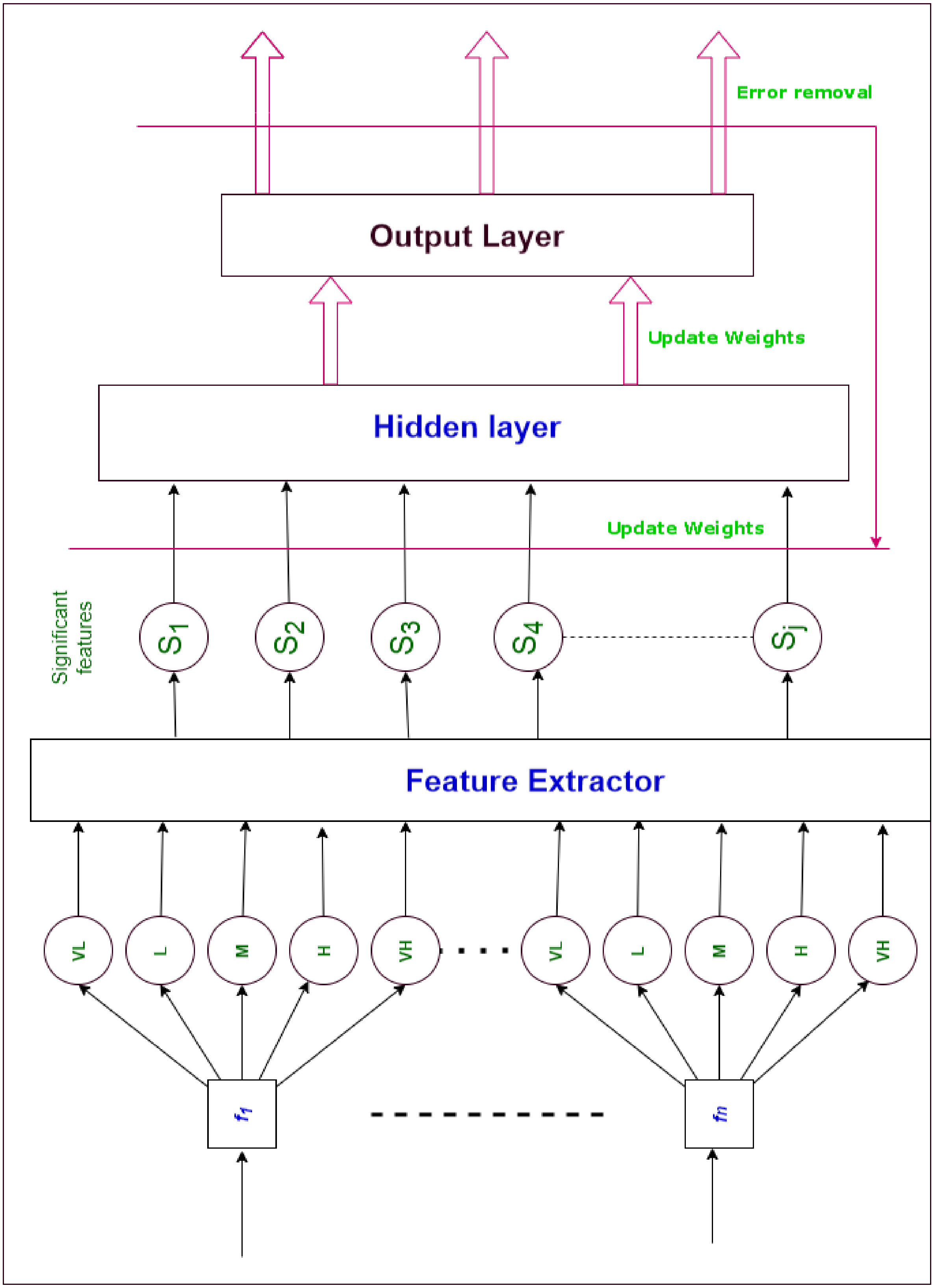

Fig. 2. Neural Network with Fuzzy Feature Vector with Various Membership Values.

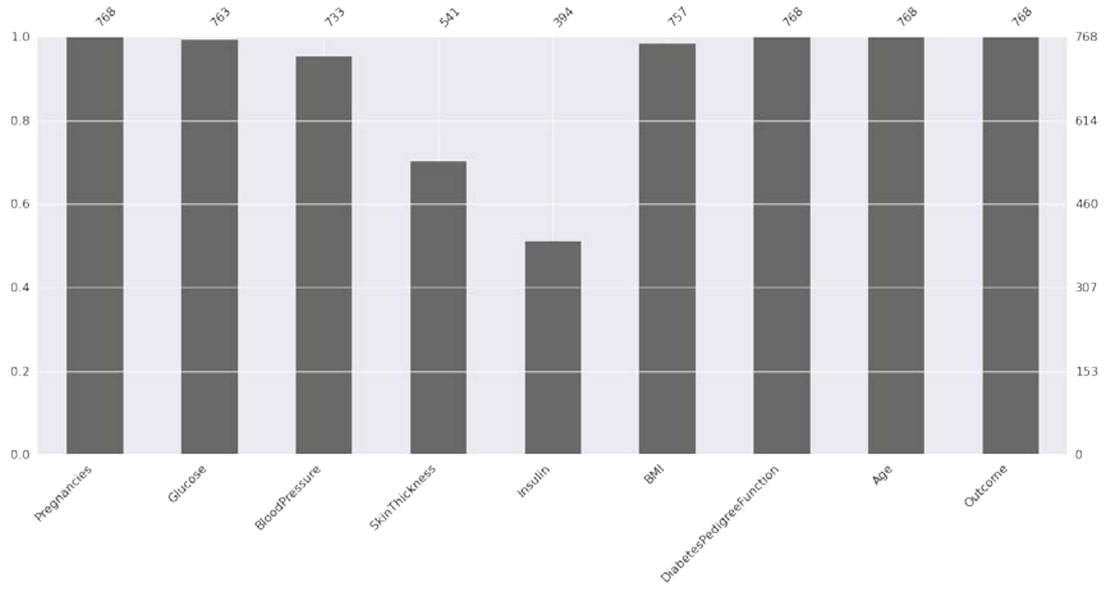

Fig. 3. PIMA Dataset.

TABLE I. PIMA DATASET DESCRIPTION

Attribute	**Attribute Description**
P	Count of number of times pregnant
G	The concentration of Plasma glucose concentration
BP	Blood pressure
ST	It gives the thickness of skin folds.
I	Two-hour serum insulin.
BMI	Body Mass Index
PF	Relatives history regarding diabetes
A	Persons age
O	Class of diabetes (Yes- 1, No -0)

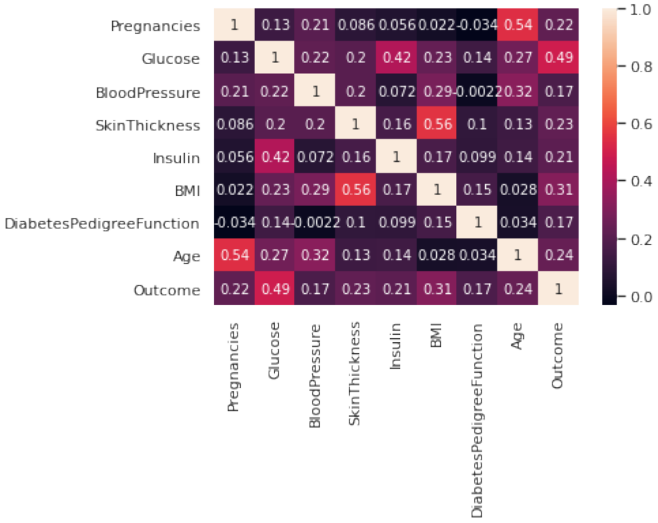

Fig. 4. Correlation of Attributes after Imputation.

Fig. 5 and 6 show the attribute values for a diabetes and non-diabetes patient. Fig. 5 shows the attributes of the PIMA dataset. The figure contains all the attribute values of the Diabetes patient.

Fig. 6 shows the attributes of a non-diabetic patient in the PIMA dataset. From both the figures it can be observed that there is a clear distinction in the ranges of the values of most of the attributes.

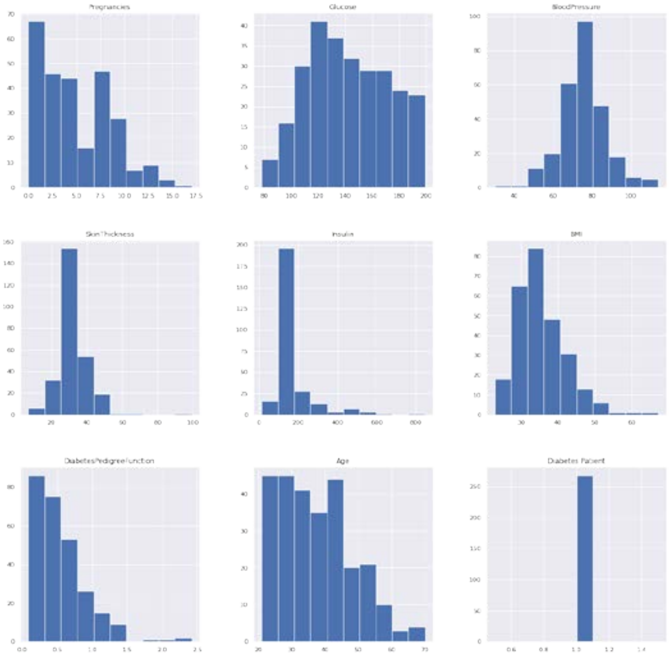

Fig. 5. Attributes for a Diabetic Patient in PIMA Dataset after Imputation.

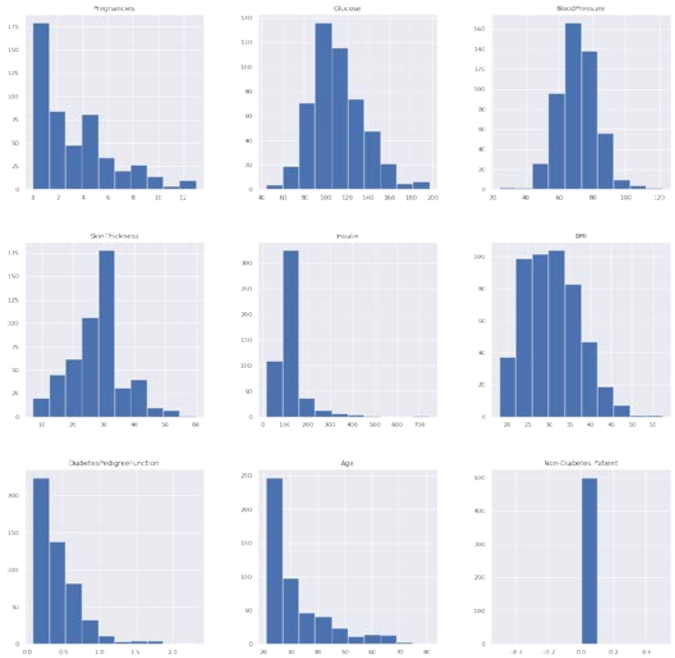

Fig. 6. Ranges of the Attributes for a non-Diabetic Patient in PIMA Dataset after Imputation.

To classify the dataset, the dataset is split into 70-30 ratio for training and testing purposes. The feature vector is the input to the neural network. The input layer assigns weights initially and passes this to the hidden layer and performs the classification task. The weights are updated until the network reaches a certain threshold. The next section provides the results of the proposed approach.

IV. Results and Discussion

In this section, the experimentation and results of the proposed approach on the PIMA dataset are given. Machine learning approaches like Decision trees, J48, SVM, Logistic regression are used for comparison.

The following are the evaluation parameters used in this paper to evaluate the performance of different models:

$$Accuracy = \frac{Number of correct predictions}{Total predictions} \quad (3)$$

$$precision = \frac{tr\_pos}{tr\_pos + fl\_pos} \quad (4)$$

$$recall = \frac{tr\_pos}{tr\_pos + fal\_neg} \quad (5)$$

$$F1 = 2 * \frac{pr * re}{pr + re} \quad (6)$$

Table II shows the results of various approaches on the PIMA dataset. From the results, it is observed that the proposed approach outperformed all the existing machine learning algorithms.

TABLE II. Comparison of different Algorithms with the Proposed Algorithm Enhanced Naïve Bayes Classifier on PID Dataset

Model	Accuracy	Precision	Recall	F1-score
Decision tree	73.16	0.73	0.73	0.70
SVM	71.85	0.71	0.72	0.68
J48	73.16	0.73	0.73	0.70
Logistic Regression	70.80	0.68	0.69	0.65
Naïve Bayes	76.92	0.77	0.77	0.77
ANN	82.00	0.81	0.81	0.80
ANN with fuzzy rules	82.33	0.85	0.83	0.81
Proposed Approach	**84.66**	**0.87**	**0.86**	**0.82**

Machine learning models like Decision tree and SVM achieved 73.16, 71.85 accuracy levels. J48 and Logistic regression model achieved 73.16 and 70.80 accuracy. Compared to these machine learning models Naïve Bayes achieved highest results. But the performance of neural networks on PIMA dataset is relatively high. The neural network model integrated with fuzzy rules also outperformed the other machine learning model results. But the highest accuracy level 84.66 is obtained when PIMA dataset is tested with the proposed model. The proposed model is able to outperform the existing models because of the following reasons:

1) The proposed model introduced fuzzy membership values for each feature in the input feature vector.

2) The significant features of the model are retained using principal component analysis.

3) The model is trained with artificial neural network and the weights are updated till there is minimal error.

The proposed work compared to other models has more advantages as the model is based on neuro fuzzy model. The model can be easily enhanced to predict the other diseases at early stages. The proposed model has outperformed all the machine learning algorithms. It is easy to draw the inferences using fuzzy models to predict the diseases and also the fuzzy model enables to easily classify the data compared to the other models.

V. Conclusion

Diabetes is considered a very serious disease in recent times since it causes many health complications. Hence, the researchers are trying to bridge the gap between technology and the medical field by developing various methods to ease the treatment procedures. In this regard, this work proposed a neural network model using fuzzy linguistic terms to classify diabetes data. The proposed model used linguistic terms to assign membership values to the features of the data and performed feature extraction using principal component analysis. After feature reduction, the features are passed to the input layer of the artificial neural network and the weights of the neural network in the input and hidden layers are updated until the error minimizes. The main goal of the proposed research is to predict the disease at early stages to deal with the other complications that may arise in the future. The proposed model aims to develop a robust approach for the early detection of diseases to mitigate the adversaries that the disease may cause to the patient and to help the practitioners in medical domain to diagnose the diabetic patients. Moreover, the proposed model is able to outperform all the existing machine learning approaches.

In future, the research can be easily extended to predict and diagnose other diseases. The proposed work can be modified with fuzzy inference rules to improve the performance. The present study can deal with the missing data of smaller datasets and in future it can be enhanced to handle large amounts of data.

References

[1] American Diabetes Association. (2019). 5. Lifestyle management: standards of medical care in diabetes—2019. *Diabetes care*, *42*(Supplement 1), S46-S60.

[2] Chen, P; and Pan, C., (2018). Diabetes classification model based on boosting algorithms. *BMC bioinformatics*, *19*(1), pp.1-9.

[3] Choubey, D. K.; Kumar, P.; Tripathi, S.; and Kumar, S. (2020). Performance evaluation of classification methods with PCA and PSO for diabetes. Network Modeling Analysis in Health Informatics and Bioinformatics, 9(1), 1-30.

[4] Maniruzzaman, M.; Rahman, M. J.; Ahammed, B.; and Abedin, M. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health information science and systems, 8(1), 1-14.

[5] Ogedengbe, M. T.; and Egbunu, C. O. (2020). CSE-DT Features selection technique for Diabetes classification. Applications of Modelling and Simulation, 4, 101-109.

[6] Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

[7] Übeyli, E. D. (2009). Adaptive neuro-fuzzy inference system for classification of ECG signals using Lyapunov exponents. *Computer methods and programs in biomedicine*, *93*(3), 313-321

[8] Maniruzzaman, M.; Kumar, N.; Abedin, M. M.; Islam, M. S.; Suri, H. S.; El-Baz, A. S.; and Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Computer methods and programs in biomedicine, 152, 23-34.

[9] Khan, N.; Gaurav, D.; and Kandl, T. (2013). Performance evaluation of Levenberg-Marquardt technique in error reduction for diabetes condition classification. Procedia Computer Science, 18, 2629-2637.

[10] Ganji, M. F.; and Abadeh, M. S. (2011). A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. Expert Systems with Applications, 38(12), 14650-14659.

[11] Perveen, S.; Shahbaz, M.; Guergachi, A.; and Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science, 82, 115-121.

[12] Mercaldo, F.; Nardone, V.; and Santone, A. (2017). Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia computer science, 112, 2519-2528.

[13] Edla, D. R.; and Cheruku, R. (2017). Diabetes-finder: a bat optimized classification system for type-2 diabetes. Procedia computer science, 115, 235-242.

[14] Sisodia, D.; and Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia computer science*, *132*, 1578-1585.

[15] G. Swapna; K. P. Soman; and R. Vinayakumar (2018). “Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals,” Procedia Comput. Sci., vol. 132, pp.1253–1262.

[16] A. Mohebbi; T. B. Aradóttir; A. R. Johansen; H. Bengtsson; M. Fraccaro and M. Mørup (2017). A deep learning approach to adherence detection for type 2 diabetics. IEEE Engineering in Medicine and Biology Society, pp. 2896–2899, 2017.

[17] T. Pham; T. Tran; D. Phung; and S. Venkatesh (2017), Predicting healthcare trajectories from medical records: A deep learning approach. J. Biomed. Inform., vol.69, pp.218–229.

[18] H. Balaji; N. Iyengar; and R. D. Caytiles(2017). Optimal Predictive analytics of Pima Diabetics using Deep Learning. Int. J. Database Theory Appl., vol. 10, pp. 47–62, 2017.

[19] Prasad, K. S.; Reddy, N. C. S.; and Puneeth, B. N. (2020). A Framework for Diagnosing Kidney Disease in Diabetes Patients Using Classification Algorithms. SN Computer Science, 1(2), 1-6.

[20] Bakris, G. L.; Agarwal, R.; Anker, S. D.; Pitt, B.; Ruilope, L. M.; Rossing, P.; and Filippatos, G. (2020). Effect of finerenone on chronic kidney disease outcomes in type 2 diabetes. New England Journal of Medicine, 383(23), 2219-2229.

[21] McGuire, D. K.; Shih, W. J.; Cosentino, F.; Charbonnel, B.; Cherney, D. Z.; Dagogo-Jack, S.; and Cannon, C. P. (2021). Association of SGLT2 inhibitors with cardiovascular and kidney outcomes in patients with type 2 diabetes: a meta-analysis. JAMA cardiology, 6(2), 148-158.