International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
IJISAE
ISSN:2147-6799 www.ijisae.org Original Research Paper

Diabetes Mellitus Disease Prediction and Classification using Latent Dirichlet Allocation and Artificial Neural Network Classifier

Soumya K N[1], Raja Praveen K N[2]

Submitted: 18/10/2023 **Revised**: 15/12/2023 **Accepted**: 22/12/2023

Abstract: The Diabetes Mellitus (DM) is known as of the persistent disease which is due to excessive blood sugar levels. When it is left untreated it leads to severe health complications like cardiac disorders, kidney damage and stroke. The existing methods based on machine learning and deep learning approaches faces problem in predicting the diabetes of the patients in a precise manner. Moreover, the classification accuracy was diminished when evaluated for large datasets, so this research introduced an effective classification approach using the combination of Latent Dirichlet Allocation (LDA) and Artificial Neural Network (ANN). The probability distribution function of LDA is combined using the back propagation of ANN where the weights are initialized to perform an effective diabetes classification. The data is obtained from PIMA Dataset and North California State University (NCSU) dataset then the pre-processing is performed using min-max normalization approach. After this, Bivariate filter based feature selection is performed to choose appropriate features that were selected using the bivariate filter method is fed as input to Pearson correlation which selects the effective features based on threshold value. Finally, classification is performed using the proposed ANN-LDA. The results show that suggested method performs better than the existing approaches and achieves classification accuracy of 93%.

Keywords*: Artificial Neural Network***,** *Bivariate filter***,** *Diabetes mellitus, Latent Dirichlet Allocation, Pearson correlation.*

1. Introduction

The pancreas relies as an important organ in the human body which secretes insulin. The insulin is responsible to maintain sugar content, fat and metabolism for production of day-to-day energy level for individuals. When the quantity of insulin is low, the glucose or sugar content in the blood gets increased. The unwanted sugar content in the blood is driven out in the form of urine which is referred as diabetes mellitus [1]. In other words, diabetes is referred as a type of metabolic disease which elevates the glucose level in the blood and leads to stroke, kidney failure and cardiovascular disease [2,3]. The type of diabetes gets varied from patient to patient based on the genetics, diagnostic criteria and region epidemics. An absolute insulin deficiency in body due to attack in pancreatic beta cells causes type 1 diabetes and the incapability of the person's body to produce insulin by their own is known as type 2 diabetes. Both the type of diabetes is increasing in a rapid manner, but the increase of type 2 diabetes is comparatively higher than the type 1 diabetes [4,5]. When the type 2 diabetes is untreated, the person body becomes insulin dependent and even leads to death [6]. So, the researchers emphasize the status of diabetes mellitus using machine learning and deep learning algorithms [7]. Moreover, the usage of these AI techniques helps to enhance the accuracy of classification and predict the type of diabetes.

The machine learning and deep learning techniques plays an important role in suggesting better decision to the doctor by predicting and classifying the type of diabetes in patients [8,9]. The machine learning and deep learning approaches aims to create computer systems on the basis of different patterns used in training the data and to perform an effective classification to predict diabetes [10,11]. Moreover, the machine learning approaches analyze the data comprised with features and the observations of the pre-labeled data which classify the diabetic classes. The volume of medical data related to diabetes is large so it is essential to differentiate the relevant features which ease the process of classification [12]. So, an effective feature selection approach minimizes the complexity during classification and aids in better classification accuracy. The classification accuracy is based on the efficiency of the classifier in detecting the patients with and without diabetes [13,14]. The usage of single classifiers results in poor classification accuracy rather than the combinational classifiers. The machine learning classifiers requires high memory space and requires more training time [15]. So, this research introduced an effective classification approach using the combination of machine learning and deep learning technqiues which classify the diabetes with better classification accuracy.

The major contributions of this research are listed as follows:

[1] *Research scholar, School of Computer Science and Engineering, JAIN (Deemed to be University) Bangalore, Karnataka, India*
Soumya.kn16@gmail.com
[2] *Department of Computer Science and Engineering, Faculty of Engineering and Technology, JAIN (Deemed-to-be University), Bengaluru, 562112, India*
rajapraveen.k.n@gmail.com

1. This research introduced an effective classification approach with the help of LDA and ANN. The probability distribution function of LDA is combined with the back propagation of ANN where the weights are initialized to perform an effective diabetes classification

2. The Bivariate filter based feature selection is performed to choose appropriate features and the features which are selected utilizing the bivariate filter method is fed as input to Pearson correlation which selects the effective features based on threshold value.

The rest of the article is organized in the following manner: The Section 2 describes the related works of this research and Section 3 describes the proposed methodology of this research. The Section 4 and Section 5 describes the results and conclusion of this overall research.

2. Related works

In this section, the recent researches based on various machine learning and deep learning algorithms for diabetes classification is described as follows:

B. Shamreen Ahamed et al [16] have introduced an effective model to predict and classify the diabetes using machine learning approaches. After acquisition of raw data, the pre-processing was performed to clean the data and the one hot encoding was used to transform the categorical values to numerical data. The data augmentation was performed to remove the excess data and the oversampling approach was employed to enhance the balance among class values. Finally, the classification was performed with the help of Light Gradient Boosting Machine (LGBM). However, the suggested approach was not robust to detect the occurrence probability of the disease. Sarita Simaiya et al [17] have introduced a multistage ensemble approach to predict and categorize the diabetes. The ensembling was performed based on two cases, in first case, Naïve Bayes classifier, K-Nearest Neighbor and J48 were used as classifier and in second case, the Random Forest and JRip classifiers are used. However, the suggested approach obtained diminished accuracy while evaluating the complex classification patterns.

Pawan Whig et al [18] have introduced low code Pycaret machine learning approach to detect and categorize diabetes. The suggested approach was the combination of classification and regression models which effectively predict the severity of the diabetic patients. The process of leveraging the power of Pycaret helps to construct an effective method which help to categorize diabetes in a precise manner. However, the suggested approach was not robust for large and diverse datasets. Daliya V. K. and T. K. Ramesh et al [19] have introduced an optimized stacking ensemble approach to predict the progression of diabetics in patients. The raw data was pre-processed using cleaning and finding distribution approach and the features were selected based on upper and lower bounds. After this, the stacking was performed among logistic regression, K-nearest neighbors, classification and regression tree, support vector machine and Naïve Bayes method. The stacking of the fore mentioned classifiers effectively analyze the condition of diabetes but the suggested approach considered the false prediction results which is a critical task while classifying life threating disease like diabetes.

Ayse Dogru et al [20] have introduced a hybrid ensemble learning approach to predict and detect the diabetes at the early stage. The super learner was on the basis of cross validation approach which predicts the results of the machine learning approach in a combined way. The suggested approach utilized four base learners with a meta learner to build a super learner. Moreover, Chi-square approach was used to perform an optimal feature selection and the hyper-parameters were fine-tuned with the help of grid search algorithm. However, usage of multiple number of machine learning algorithms as a base learner enhance the computational complexity. Marwan Al-Tawil et al [21] have introduced a bio-inspired machine learning approach to detect the type 2 diabetes. In the suggested research, the cuttlefish algorithm was used in the process of extracting the significant features. After the stage of extracting the features, the output is subjected to evaluate the performance of different classifiers such as Logistic Regression, Support Vector Machine and Naïve Bayes. However, the cuttlefish optimization algorithm does not consider the computation cost factor which diminish the approximation of fitness function.

Victor Chang et al [22] have introduced a e-diagnosis system using various machine learning approaches to predict and diagnose diabetes mellitus. The ML technqiues such as Naïve Bayes, Random Forest and Decision Tree are utilized to perform an effective categorization. Among three machine learning approaches, Naïve Bayes model is fine-tuned with effective features and aids in better classification results. Jobeda Jamal Khanam and Simon Y. Foo [23] have utilized various machine learning approaches to predict the diabetes from PIMA dataset. After the stage of data acquisition, pre-processing was performed to remove outliers and regularizing the values. Then the data is trained using different machine learning approaches and classification takes place. The performance of the utilized machine learning classifiers affected when the dataset values get randomized.

In overall, the existing approaches faced problems in providing better classification accuracy due to the computational complexity, false prediction rate at complex regions. Moreover, the usage of conventional classification approaches based on deep learning and machine learning faced problems based on higher training time and memory space.

3. Diabetes Classification using ANN-LDA

In this section, the process involved in classification of diabetes using the proposed approach is explained in detail along with the processing stages. The process involved in various stages of classifying diabetes are data acquisition, pre-processing, feature selection and classification of diabetes. Initially, the data is obtained from one of the publicly available dataset and the pre-processing is performed to remove the irrelevant or inappropriate features. After this stage, the process of selecting features takes place to select the relevant and irredundant features. At last, an effective classification is performed with the help of proposed classifier. The figure 1 depicted below provides the diagrammatical representation for overall process involved in classification of diabetes.

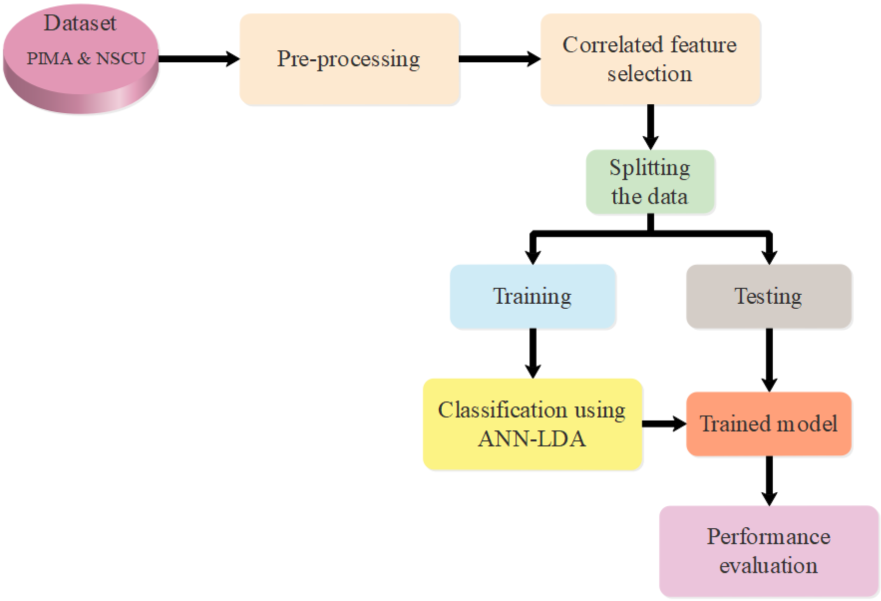


Fig 1. Overall process involved in classification of diabetes

3.1 Data acquisition

In this research, the raw data is obtained from two publicly available datasets such as PIMA Dataset [24] and North California State University (NCSU) [25] dataset. The description of those fore mentioned dataset is mentioned as follows:

PIMA: It is one of the commonly used dataset which was used by more number of researchers to work on diabetes mellitus. PIMA dataset is available in UCI repository database and it is comprised with nine attributes based on pregnancy, glucose, BP, thickness of the skin, insulin, BMI, age group and diabetes pedigree function. By including those nine attributes, 768 instances are present in PIMA dataset.

NCSU: The data from NCSU dataset is obtained from NC state university which consist of 442 instance with around 10 attributes. The feature set of NSCU is comprised with age, sex, BMI, and blood glucose level.

3.2 Data pre-processing

After the stage of data acquisition, the pre-processing is performed to transform the data to processed data without complexities. In this research, the pre-processing is performed with the help of min-max normalization approach. The min-max normalization approach normalizes the variables to even scale which ranges from 0 to 1 which is presented in equation (1) as follows:

$$x^* = \frac{\text{X}-\min(x)}{\max(x)-\min(x)}$$

(1)

Where normalized value is represented as x^* and actual value is denoted as X. Minimalized value present in the dataset is represented as min(x) and the highest dataset value is represented as max(x).

3.3 Feature selection

The pre-processed output is obtained from the stage of pre-processing which is fed as input to the stage of feature selection. The feature selection is performed to enhance classification accuracy by selecting appropriate features that ease the process of classifying diabetes. In this research, Bivariate statistics approach is used in the process of feature selection which select a relevant and appropriate feature from the large dataset like PIMA and NCSU. The feature extraction which is performed using Bivariate filter combines the heterogenous data layers to resolve the uncertain input data. Moreover, the Bivariate filter utilize a certainty factor to mine the relevant features and it is evaluated using the equation (2) as follows:

$$CF = \begin{cases} \frac{PP_a - PP_s}{PP_a(1-PP_s)}, if\ PP_a \geq PP_s \\ \frac{PP_a - PP_s}{PP_s(1-PP_a)}, if\ \ |\ PP_a < PP_s \end{cases} \quad (2)$$

Where the conditional probability of CF is denoted as PP_a and the prior probability of the selected features are represented as PP_s. The value of PP_a and PP_s is evaluated using the equation (3) and (4) respectively.

$$PP_a = P\{S|B\} \quad (3)$$

$$PP_s = P\{S\} \quad (4)$$

Where the conditional probability unit of B is represented as $P\{S|B\}$. The positive results show the increase of certainty value and the negative value denotes the decrease of certainty value. Moreover, Weight of Evidence (WoE) on the basis of Bayesian probability approach are used to extract the features based on the weights. The two parameters such as W^+ and W^- are used to evaluate the positive and negative weights of the features which is represented in equation (5) and (6) respectively.

$$W^+ = ln\frac{P\{B|A\}}{P\{B|\bar{A}\}} \ ... \quad (5)$$

$$W^- = ln\frac{P\{\bar{B}|A\}}{P\{\bar{B}|\bar{A}\}} \ ... \quad (6)$$

Where the logarithm and the probability values are represented as P and ln. The features which are selected using the bivariate filter method is fed as input to Pearson correlation which selects the effective features based on threshold value.

3.3.1 Pearson Correlation

The relation among the Pearson correlation and diabetic characteristics are improved to determine the parameters to eliminate the redundant information. The random variables in a linear relationship is evaluated using the Pearson correlation co-efficient. The equation (7) depicted below shows the linear correlation among two continuous variables.

$$r_{xy} = \frac{\Sigma(x_i - \bar{x})\ \Sigma y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2}\ \sqrt{\Sigma(y_i - \bar{y})}^2} \quad (7)$$

If $r_{xy} = 1$, x and y are a totally positive correlation, If $r_{xy} = 0$, the linear correlation between x and y is not obvious and when $r_{xy} = -1$, x and y are a totally negative correlation.

With a Pearson R-value, Pearson correlation has a less significant effect on Diabetes. With a Pearson R-value of 0.12. The relationship between feature and diabetes is moderate (r = 0.33, r = -0.42, r = 0.23). This is merely a correlation, not a proof of causation. With the aid of this one, it was possible to pinpoint the features that are highly connected and pertinent for usage as input variables in precise diabetes disease. The output of Pearson correlation is given to the classification for identifying type II diabetes disease.

3.4 Classification using ANN-LDA

After the stage of feature selection, the classification is performed with the help of PIMA and NSCU dataset. This research proposed an effective classification approach utilizing Latent Dirichlet Allocation (LDA) and Artificial Neural Network (ANN). The overall process involved in proposed ANN-LDA is deliberated in this section:

3.4.1 LDA

The LDA could learn from the hidden topic information in the form of probability distribution and LDA acts as an effective tool in modeling the topic and for the domain related to clustering. LDA topic model is a three-layered Bayesian probability model which is comprised with word, text, and topics. The relation among the document and the topic follows Dirichlet distribution and the relation among the word follows polynomial distribution. The generative process involved in LDA is represented in figure 2. The formula to compute the probability distribution function is represented in equation (8) as follows:

$$p(\theta, z, w|\alpha \text{ and } \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\,\theta) p(w_n|z_n, \beta) \quad (8)$$

where total count of samples are as M and the hidden samples are denoted as K. Layered parameters are represented as α and β, where the relative strength of the hidden latent are represented as α and the probability distribution is represented as β. The probability distribution for the specified sample is denoted as θ and the probability distribution of the hidden sample is denoted as φ. The process of sampling is represented in rectangle and the observable variables are represented in the form of bi-circle.

The overall iterative process of LDA is listed below,

- The hidden topic and the feature words are evaluated using the polynomial distribution function and Dirichlet distribution function.
- Then evaluate the probability distribution for each sample of every individual sample set.
- The hidden sample is selected in a randomized manner from the probability distribution function and the feature word is selected from the polynomial distribution function z.

3.4.2 Combining the ANN with LDA

The probability distribution function of the LDA is combined with back propagation training of ANN to offer better classification results. In back propagation of ANN, the weights are initialized as a randomized number for every individual unit associated with it. First, input layer of network receives training tuple. Inputs are unchanged as they move over the units of the input. After this, hidden and output layers, individual units, net input and output are calculated. Every individual input is evaluated when it is multiplied by its appropriate weightiness are added to control input to unit that is represented in equation (9) as follows:

$$I_j = \sum_i W_{ij}\, O_i + \theta_j \tag{9}$$

Where the weighted connection from the previous layer is represented as W_{ij}, the output and the bias of the unit is represented as θ_j. The threshold acts as a biased value which acts as varying unit and every unit considers the net input and the activation function. The sigmoid function is utilized as output unit j, is calculated in equation (10) as follows:

$$O_j = \frac{1}{1+e^{-I}j} \tag{10}$$

Where the output unit is represented as j and it is also known as squashing function since it map large input ranges from 0 to 1. The weights and biases of the hidden layer in the net input layer propagates the error and it is combined with probability distribution function of the LDA to provide an effective classification which is represented in equation (11). The equation 11 formulated from the combination of ANN and LDA aids in better classification of diabetes.

$$W_{ij} = W_{ij} + (l)error_j O_i + p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\,\theta) p(w_n|z_n, \beta) \tag{11}$$

Where the variable l is known as the learning rate which helps in effective classification and the combination of probability distribution function helps to minimize the learning rate and helps to achieve better classification accuracy.

4. Results and Analysis

Here, the results obtained from proposed ANN-LDA is evaluated to obtain the results based on diabetes classification. The result section is sub-sectioned to performance analysis and the comparative analysis. In performance analysis, the efficiency of suggested method is calculated for two different datasets such as PIMA and NSCU. In comparative analysis, the efficiency of proposed approach is estimated with existing approaches listed in related works. The metrics like accuracy, precision, recall and f-measure are measured to calculate effectiveness of suggested method which are evaluated using the equation (12-15) listed as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{12}$$

$$Precision = \frac{TP}{TP+FP} \tag{13}$$

$$Recall = \frac{TP}{TP+FN} \tag{14}$$

$$F1\, measure = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{15}$$

Where the TP, FP, TN and FN is the true positive, false positive, true negative and false negative correspondingly.

4.1 Experimental setup

The suggested approach is simulated in Python 3.7 software and implementation was done in a system with specifications such as Windows 10 OS, Intel core i7 processor and 16 GB of random access memory.

4.2 Performance analysis

In this sub-section, the effectiveness of the proposed approach is evaluated with various classifiers such as K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM) and Artificial Neural Network (ANN). The performance of the fore mentioned classifiers are evaluated based on two datasets which is represented in Table 1 and Table 2 respectively. The Table 1 depicted below shows the results which is obtained from the proposed approach for PIMA dataset.

Table 1. Comparing the performance of the classifiers for PIMA dataset

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
KNN	76	72	73	72
LR	78	74	77	75
DT	76	74	74	74
SVM	87	73	85	78
ANN	82	77	86	81
ANN-LDA	93	81	88	84

The results from the table 1 shows that the proposed ANN-LDA acts as an excellent classifier while classifying the diabetic patients from PIMA dataset. The proposed classification approach has obtained better result in overall metrics when compared with existing classification methods. For instance, the classification accuracy of the proposed ANN-LDA is 93% which is comparatively higher than the existing KNN, LR, DT, SVM and ANN with accuracies of 76%,78%,76%,87% and 82% respectively. The figure 2 presents the graphical representation of different classifier for PIMA dataset and ROC curve of ANN-LDA for accuracy value of 93% in PIMA dataset is represented in figure 3.

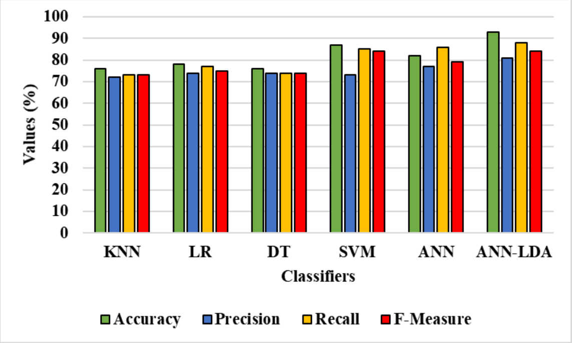


Fig 2. Graphical representation for the performance of the classifiers for PIMA dataset

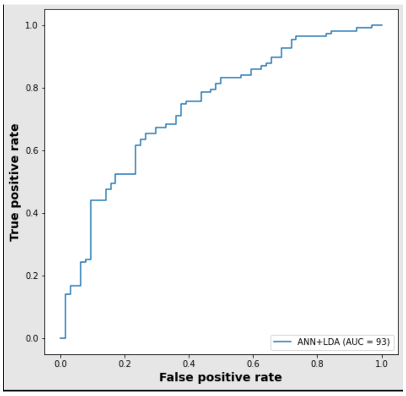


Fig 3. ROC graph of accuracy for PIMA dataset

Secondly, the performance of the proposed classifier with existing ones are evaluated for NSCU dataset. The results obtained while evaluating the suggested approach with existing classification approaches are evaluated for NSCU dataset and the obtained results are tabulated in Table 2 as follows:

Table 2. Comparing the performance of the classifiers for NSCU dataset

Classifiers	**Accuracy(%)**	**Precision (%)**	**Recall (%)**	**F-Measure(%)**
KNN	82	79	86	82
LR	75	70	73	71
DT	80	80	83	81
SVM	85	77	77	77
ANN	78	75	73	73
ANN-LDA	**93**	**81**	**82**	**82**

The results from the table 2 shows that the proposed classification approach have achieved better results in overall performance metrics. For example, the classification accuracy of the proposed ANN-LDA for NSCU dataset is 93% whereas the classification accuracy of the existing KNN, LR, DT, SVM and ANN is 82%,75%,80%,85%, and 78% respectively. These results show the efficiency of the proposed approach while classifying the diabetic patients in NSCU dataset. The graphical representation for the performance of the classifiers for NSCU dataset is described in figure 4 and ROC curve of ANN-LDA with accuracy value of 93% for NSCU dataset is represented in figure 5.

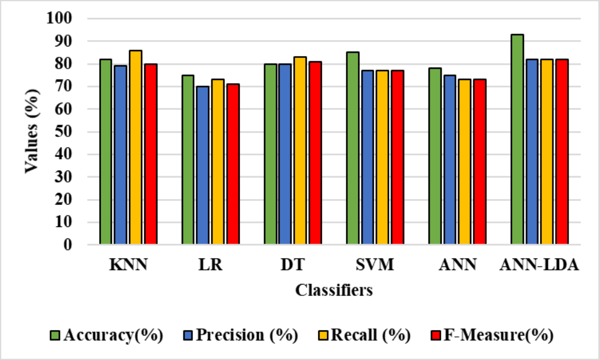


Fig 4. Graphical representation for the performance of the classifiers for NSCU dataset

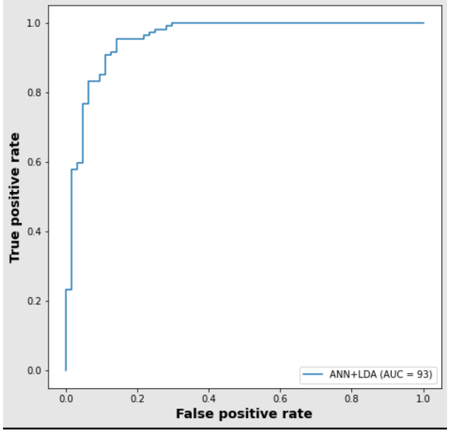


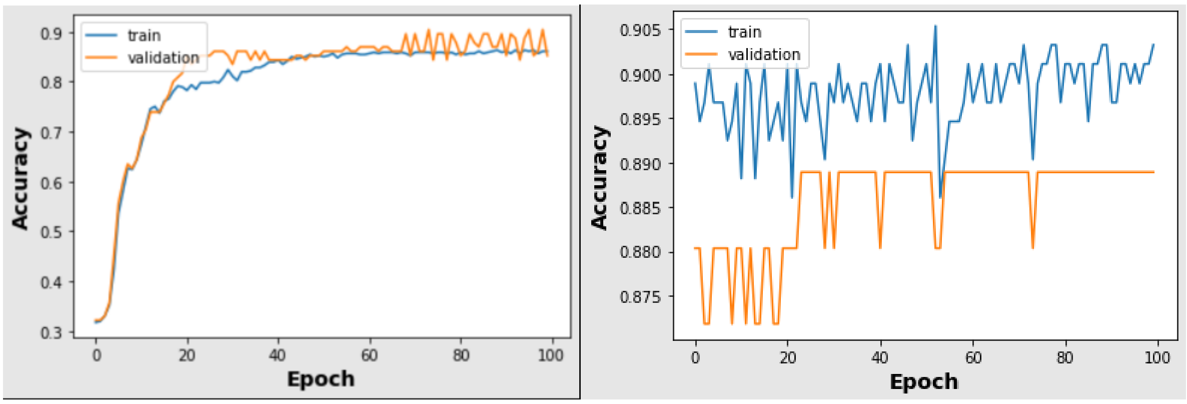
Fig 5. ROC graph of accuracy for NSCU dataset

Thirdly, the performance of the proposed ANN-LDA classifier is evaluated for different threshold values for selecting the appropriate features. The table 3 depicted below presents the value of accuracy, precision, recall and F-1 score for different threshold value of 0,0.1,0.2,0.3 and 0.4.

Table 3. Evaluating the performance of ANN-LDA for different threshold values

Threshold value	Accuracy(%)	Precision (%)	Recall (%)	F1-score(%)
0	89	81	89	84
0.1	88	81	89	84
0.2	88	78	85	81
0.3	87	77	92	83
0.4	84	74	79	76

The accuracy obtained while evaluating the proposed classifier based on training and validation for NCSU dataset is presented in 6(a) and for PIMA dataset is presented in 6(b).



6(a)

6(b)

Fig 6. Training and validation accuracy for 100 epochs for NCSU dataset and PIMA dataset

4.3 Comparative analysis

In this subsection, the performance of the proposed classification approach is evaluated with existing approaches listed in related works. The performance is evaluated based on performance metrics such as accuracy, precision, recall and F-1 score. The results obtained while evaluating the proposed approach for PIMA dataset is listed in table 3 as follows:

Table 3. Comparing the classification efficiency of the proposed approach with existing approach for PIMA dataset

Classifiers	Accuracy(%)	Precision (%)	Recall (%)	F-Measure(%)
Multi-stage ensemble [17]	-	78.4	78.6	78.5
Naïve Bayes [22]	77	73	75	74
LR [23]	78	78	78	78
ANN-LDA	**93**	**81**	**88**	**84**

The results from the table 3 shows that the proposed classification approach achieved better results in overall performance metrics. For example, the precision is considered as a common performance metric to evaluate the efficacy of the suggested approach. The precision of the proposed approach is 81% which is comparatively higher than the existing Multi-stage ensemble method, Naïve Bayes and LR with value of 78.4%, 73% and 78% respectively.

5. Conclusion

This research proposed an effective classification approach using LDA-ANN.The probability distribution function of LDA is combined with the back propagation of ANN where the weights are initialized to perform an effective diabetes classification. The performance of the proposed approach is evaluated for PIMA and NCSU datasets based on accuracy, precision, recall and F-measure. After the stage of pre-processing using min-max normalization, Bivariate filter based feature selection is performed to select the relevant features and the features which are selected using the bivariate filter method is fed as input to Pearson correlation which selects the effective features based on threshold value. Finally, classification is performed using the proposed ANN-LDA. The results show that the proposed approach achieved better results in overall metrics. The classification accuracy of the proposed approach is 93% which is comparatively higher than the existing Naïve Bayes and LR with value of 77% and 78% respectively. In future, the combination of meta-heuristic algorithms can be included to select the appropriate features which improve the classification accuracy.

References

[1] Chang, V., Bailey, J., Xu, Q.A. and Sun, Z., 2022. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, pp.1-17.

[2] aiswal, S. and Gupta, P., 2023. Diabetes Prediction Using Bi-directional Long Short-Term Memory. *SN Computer Science*, *4*(4), p.373..

[3] Yuan, Z., Ding, H., Chao, G., Song, M., Wang, L., Ding, W. and Chu, D., 2023. A Diabetes Prediction System Based on Incomplete Fused Data Sources. *Machine Learning and Knowledge Extraction*, *5*(2), pp.384-399..

[4] Alhmiedat, Tareq, and Mohammed Alotaibi. "The Investigation of Employing Supervised Machine Learning Models to Predict Type 2 Diabetes Among Adults." KSII Transactions on Internet & Information Systems 16, no. 9 (2022).

[5] Lu, H., Uddin, S., Hajati, F., Moni, M.A. and Khushi, M., 2022. A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus. *Applied Intelligence*, *52*(3), pp.2411-2422.

[6] Alex, S.A., Nayahi, J.J.V., Shine, H. and Gopirekha, V., 2022. Deep convolutional neural network for diabetes mellitus prediction. *Neural Computing and Applications*, *34*(2), pp.1319-1327.

[7] Phongying, M. and Hiriote, S., 2023. Diabetes Classification Using Machine Learning Techniques. *Computation*, *11*(5), p.96.

[8] Butt, U.M., Letchmunan, S., Ali, M., Hassan, F.H., Baqir, A. and Sherazi, H.H.R., 2021. Machine learning based diabetes classification and prediction for healthcare applications. *Journal of healthcare engineering*, *2021*.

[9] Shafi, S. and Ansari, G.A., 2021, May. Early prediction of diabetes disease & classification of algorithms using machine learning approach. In *Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*.

[10] Sonia, J.J., Jayachandran, P., Md, A.Q., Mohan, S., Sivaraman, A.K. and Tee, K.F., 2023. Machine-Learning-Based Diabetes Mellitus Risk Prediction Using Multi-Layer Neural Network No-Prop Algorithm. *Diagnostics*, *13*(4), p.723.

[11] Chauhan, A.S., Varre, M.S., Izuora, K., Trabia, M.B. and Dufek, J.S., 2023. Prediction of Diabetes Mellitus Progression Using Supervised Machine Learning. *Sensors*, *23*(10), p.4658.

[12] Kumar, A., Kaur, A., Singh, P., Driss, M. and Boulila, W., 2023. Efficient Multiclass Classification Using Feature Selection in High-Dimensional Datasets. *Electronics*, *12*(10), p.2290.

[13] Tasin, I., Nabil, T.U., Islam, S. and Khan, R., 2023. Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, *10*(1-2), pp.1-10.

[14] Uddin, M.A., Islam, M.M., Talukder, M.A., Hossain, M.A.A., Akhter, A., Aryal, S. and Muntaha, M., 2023. Machine Learning Based Diabetes Detection Model for False Negative Reduction. *Biomedical Materials & Devices*, pp.1-17.

[15] García-Domínguez, A., Galván-Tejada, C.E., Magallanes-Quintanar, R., Gamboa-Rosales, H., Curiel, I.G., Peralta-Romero, J. and Cruz, M., 2023. Diabetes Detection Models in Mexican Patients by Combining Machine Learning Algorithms and Feature Selection Techniques for Clinical and Paraclinical Attributes: A Comparative Evaluation. *Journal of Diabetes Research*, *2023*.

[16] Ahamed, B.S., Arya, M.S., Sangeetha, S.K.B. and Auxilia Osvin, N.V., 2022. Diabetes Mellitus Disease Prediction and Type Classification Involving Predictive Modeling Using Machine Learning Techniques and Classifiers. *Applied Computational Intelligence and Soft Computing*, *2022*

[17] Simaiya, S., Kaur, R., Sandhu, J.K., Alsafyani, M., Alroobaea, R., Margala, M. and Chakrabarti, P., 2022. A novel multistage ensemble approach for prediction and classification of diabetes. *Frontiers in Physiology*, *13*, p.1085240

[18] Whig, P., Gupta, K., Jiwani, N., Jupalle, H., Kouser, S. and Alam, N., 2023. A novel method for diabetes classification and prediction with Pycaret. *Microsystem Technologies*, pp.1-9

[19] VK, D. and Ramesh, T.K., 2023. Optimized stacking ensemble models for the prediction of diabetic progression. *Multimedia Tools and Applications*, pp.1-25.

[20] Doğru, A., Buyrukoğlu, S. and Arı, M., 2023. A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. *Medical & Biological Engineering & Computing*, *61*(3), pp.785-797.

[21] Al-Tawil, M., Mahafzah, B.A., Al Tawil, A. and Aljarah, I., 2023. Bio-Inspired Machine Learning Approach to Type 2 Diabetes Detection. *Symmetry*, *15*(3), p.764.

[22] Chang, Victor, Jozeene Bailey, Qianwen Ariel Xu, and Zhili Sun. "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms." Neural Computing and Applications (2022): 1-17.

[23] Khanam, Jobeda Jamal, and Simon Y. Foo. "A comparison of machine learning algorithms for diabetes prediction." Ict Express 7, no. 4 (2021): 432-439.

[24] Link for PIMA dataset: https://www.kaggle.com/uciml/pima-indians-diabetes-database

[25] SVKR Rajeswari, Vijayakumar Ponnusamy. "Prediction of diabetes mellitus using machine learning algorithm." Annals of the Romanian Society for Cell Biology (2021): 5655-5662.