**Journal of Artificial Intelligence, Machine Learning and Neural Network**
**ISSN:2799-1172**
Vol: 02, No. 04, June-July 2022
http://journal.hmjournals.com/index.php/JAIMLNN
**DOI:** https://doi.org/10.55529/jaimlnn.241.8

# Artificial Neural Networks Based Predictive Model for Detecting the Early-Stage Diabetes

**Shokhjakhon Abdufattokhov[1*], Nodira Normatova[2], Makhbuba Shermatova[3]**

*[1*,2]Automatic Control and Computer Engineering Department, Turin Polytechnic University in Tashkent, 17 Little Ring Road, Tashkent, Uzbekistan*
*[3]Automatic Mathematical Natural Sciences Department, Turin Polytechnic University in Tashkent, 17 Little Ring Road, Tashkent, Uzbekistan*

*Email: [2]n.normatova@ polito.uz, [3]m.shermatova@ polito.uz*
*Corresponding Email: [1*]sh.abdufattohov@polito.u*

**Received:** 20 February 2022 **Accepted:** 10 May 2022 **Published:** 15 June 2022

***Abstract: High blood glucose levels cause diabetes, and it is characterized as a chronic disease that will disrupt fat and protein metabolism. The blood glucose levels rise because it cannot be burned in the cells due to a shortage of insulin secretion by the pancreas, or the insulin produced by the cell is insufficient. If exact early detection is possible, the hazard and prevalence of diabetes can be decreased considerably. With this, the application of technology has been an essential part of providing accurate and acceptable results in the prevention and early detection of the illness. This research implements artificial neural networks to predict the early stage of diabetes by incorporating methods involving feature selection or dimension reduction using a Relief-Based Filter for testing and training data. The results show 99.3% prediction accuracy and can be essential in contributing to a new way that is highly accurate in determining diabetes among patients.***

***Keywords: Artificial Neural Networks, Blood Glucose, Diabetes, Relief-Based Filter.***

## 1. INTRODUCTION

High blood glucose levels cause diabetes, and it is characterized as a chronic disease and also will disruption of fat and protein metabolism. The blood glucose levels rise because they cannot be burned in the cells due to a lack of insulin secretion by the pancreas or insufficient insulin produced by the cell [1]. Many people are not fully aware of having early-stage diabetes symptoms. It causes frequent urination and intensifies thirst and hunger. Diabetes is influenced by various factors such as height, weight, genetic factors, and insulin, but the most crucial factor to remember is sugar concentration. The only way to avoid problems is to detect their problem early on. The suitable classifier method and important data used were significant to classify the early stage of diabetes. Other researchers used machine learning classification to diagnose diseases. The machine learning researchers used are Random Forest, Decision Tree, Naive Bayes, Support Vector Machines, etc. As a result of their study, machine learning algorithms are more effective in diagnosing diseases [3],[4],[5].

Machine learning and data mining handle extensive data from various sources, and study context information is incorporated [6],[7],[8]. To determine the prediction of diabetes in patients, an artificial neural network learning model is used and tested on the Early-Stage Diabetes dataset. However, implementing machine learning algorithms to detect the Early-Stage Diabetes Dataset is not new. The accuracy was achieved using Random Forest Algorithm applying a tenfold cross-validation and percentage split technique with 97.4% without using feature selection [10]. Researchers conclude that the performance of all three algorithms is assessed using various metrics such as precision, accuracy, F-Measure, and recall. In comparison to other algorithms, the results show that Nave Bayes has the highest accuracy of 76.30 % [11]. Based on the findings of Alam et al., 2019 diabetes was predicted using artificial neural networks (ANN), random forests (RF), and K-means clustering techniques. The ANN technique had the highest accuracy of 75.7 %, and could help assist medical professionals with treatment decisions [12]. Random forest and decision tree have the highest specificity percentage of 98.00% and 98.20% for the classification of diabetic data. Additionally, naive Bayesian has the highest percentage in terms of accuracy with 85.30% to improve classification accuracy. The study generalizes features selection from a dataset [15]. Based on the Kandhasamy & Balamurali random forest study, KNN has the highest accuracy provided 100%. Researchers conclude that removing noisy data greatly affects generating the best result. For future studies, the said proposed method will be able to use for any disease study a suitable dataset [16]. In addition, diabetes risk classification is based on patient symptom information using ANN. For neural network training, the scaled conjugate gradient backpropagation technique was used. The classification system was 99.2% accurate in predicting the likelihood of diabetes [17].

## 2. RESEARCH METHODOLOGY

This section shows the research workflow explaining the details of the proposed methodology, as shown in Figure 1. The dataset considered in this paper contains 520 persons with diabetes-related symptoms, including data that peoples have symptoms that may cause diabetes. Dataset has an attribute of 16 as described in Table 1 and 750 instances.

Table 1: Description of Dataset Features.

| **Attributes** | **Values** | |
|---|---|---|
| Class | Positive - 1 | Negative - 0 |
| Obesity | Yes - 1 | No - 0 |
| Alopecia | Yes - 1 | No - 0 |
| Muscle stiffness | Yes - 1 | No - 0 |
| Partial paresis | Yes - 1 | No - 0 |
| Delayed healing | Yes - 1 | No - 0 |
| Irritability | Yes - 1 | No - 0 |
| Itching | Yes - 1 | No - 0 |

| | | |
|---|---|---|
| Visual blurring | Yes - 1 | No - 0 |
| Genital thrush | Yes - 1 | No - 0 |
| Polyphagia | Yes - 1 | No - 0 |
| Weakness | Yes - 1 | No - 0 |
| Sudden weight loss | Yes - 1 | No - 0 |
| Polydipsia | Yes - 1 | No - 0 |
| Polyuria | Yes - 1 | No - 0 |
| Sex | Male - 1 | Female - 0 |
| Age | 1 - (20-35), 2 – (36-45), 3 – (46-55), 4 – (56-65), 5 - (above 65) | |

### A. Feature Selection

The nearest neighbors in the space are identified it's all characteristics via ReliefF features selection model. Equation (1) the distance between instances $R_i$ and $R_j$ is calculated in the space of all attributes $\alpha \in A$, generally using a Manhattan (q = 1) metric, but it may alternatively be calculated using a Euclidean (q = 2) metric

$$D_{ij=}(\textstyle\sum_{\alpha\in A}|diff(a(R_i,R_j))|q)^{1}/_{q}, \qquad (1)$$

in which the usual $\mathcal{L}$ shown in equation (2) formula for a real-valued attribute a between two instances $R_i$ and $R_j$ is

$$\mathcal{L}\left(a(R_i,R_j)\right)=\frac{|value(a,R_i)-value(a,R_j)|}{max(a)-min(a)} \qquad (2)$$

This $\mathcal{L}$ is suitable for gene expression and other genuine predictors. For genome-wide association study (GWAS) data with categorical characteristics, just modify the $\mathcal{L}$, but the algorithm remains unchanged. The $\mathcal{L}$ function is needed by Relief techniques to construct the distance matrix for locating nearest hit and miss neighbors, but it is also required to compute the Relief significance scores [34].

### B. Artificial Neural Networks

Three layers of the artificial neural network are made up of an input layer of neurons or nodes one, two or more three hidden layers of neurons, units, the final layer is output neurons. Simple architecture with two hidden layers and weighting lines connecting neurons is shown in Figure 2.

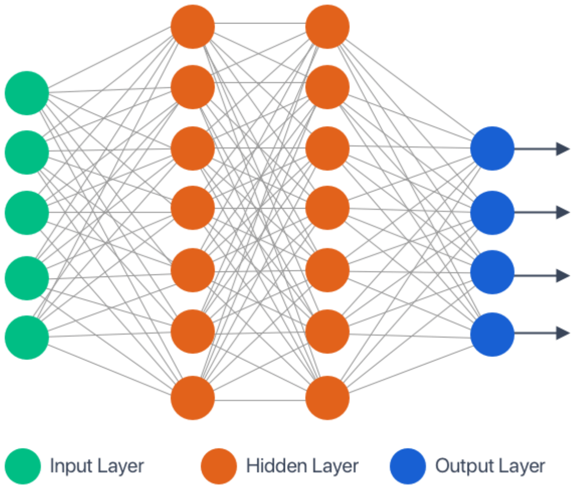


Fig.2 Artificial Neural Network Architecture.

In ANN design, weights $W$ is assigned in each connection, which is a numeral value $h_i$ is the output of neuron *I* in the hidden layer

$$h_i = z\left(\sum_{j=1}^{N} W_{ij}\, x_j + b_i^{hid}\right) \qquad (3)$$

through an activation or transfer function $z(\bullet)$, the number of input neurons is $n$, $W_{ij}$ are the weights, $x_j$ are the inputs to the input neurons, and $b_i^{hid}$are the hidden neurons' bias terms. In addition to providing nonlinearity into the neural network, the persistence of the activation function is to make the value of the neuron so that split neurons do not freeze the neural network. The common example of an activation function is the sigmoid or logistic function $g(z)$ defined as

$$g(z) = \frac{1}{1 + \exp(-z)} \qquad (4)$$

whose graphical representation is given in Figure 3 together with alternative commonly used activation functions.

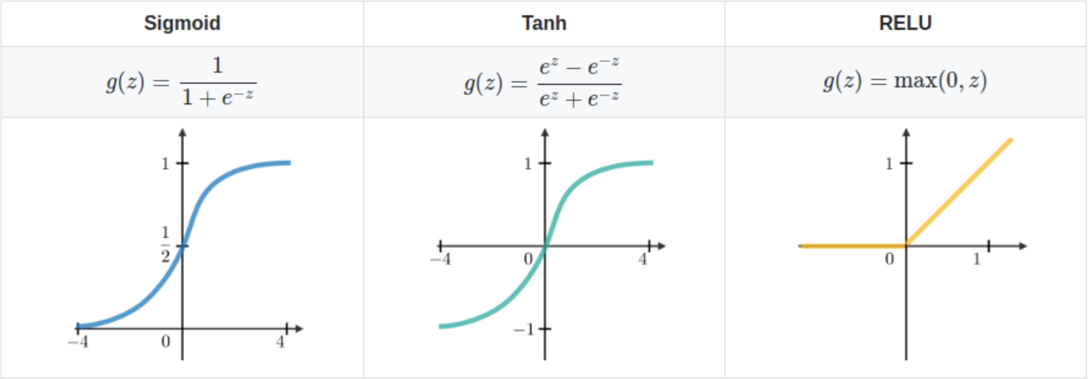


Fig.3 Common Activation Functions for ANN.

## C. Confusion Matrix

The confusion matrix gives a matrix output and describes the model's overall performance as given in Table 2

Table 2: Performance Measure.

| Measure | Derivations |
|---|---|
| Accuracy | TP + TN / TP + TN + FN + FP |
| Precision | TP / FP + TP |
| Recall | TP / FN + TP |
| Specificity | TN / TN + FP |

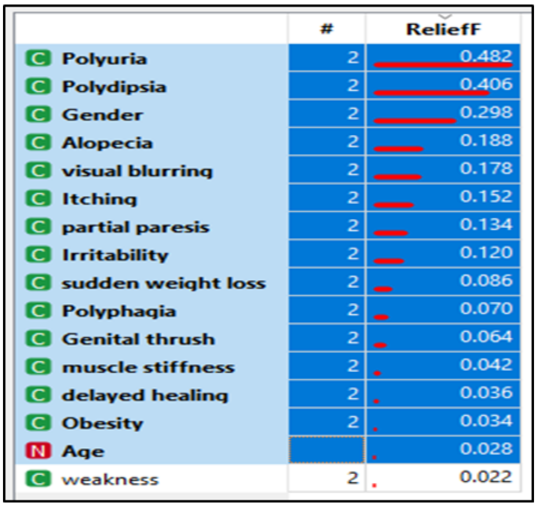
where

- True Positive (TP) - predicted the positive actual values of a dataset,
- False Positive (FP) - an error rate of predicted values for incorrect predicted the positive actual values,
- False Negative (FN) - an error rate of predicted values for incorrect predicted the actual negative values and
- True Negative (TN) - predicted the actual negative values of the dataset.

Evaluate the performance of each model using the elements of the confusion matrix. The classification models' accuracy, precision, recall, and specificity may be calculated. Several measures for evaluating the performance of different categorization methods may be extracted from the confusion matrix. Consideration of classification accuracy or its balance error rate is a popular assessment technique. Several more metrics are calculated and compared based on the values in the confusion matrix. The algorithm's accuracy measures the likelihood of predicting positive and negative records adequately. Precision is computed by dividing the number of correct positive outcomes by the number of positive outcomes predicted by the classifier. Precision is defined as the proportion of relevant results in the list of all returned search results. The probability of correctly predicting negative cases is referred to as specificity [39].

## 3. RESULTS

We used the ReliefF features selection method for dimensional reduction to select the best attributes in our experimentation and rank all attributes using a surrogate measure of feature (value), followed by defining the feature subset using an arbitrary cut off. This cut off can be determined by the predictive or subjective likelihood of relevance or simply by the number of features in the subset desired. The following are the result of all attributes with defined values as summarized in Figure 4. Based on the result, the 'Weakness' attribute had the lowest subjective likelihood significance having 0.022. Furthermore, the 'Weakness' attribute is no longer one of the attributes utilized in the dataset because of unlikelihood symptoms for the early stage of diabetes. Additionally, it affects the result of accuracy in getting high accuracy percentage in the machine learning model. As a result, from the initial dataset of 16 attributes, 15 attributes and 468 instances are utilized by applying features selection for dimensional reduction.



| | | # | ReliefF |
|---|---|---|---|
| C | Polyuria | 2 | 0.482 |
| C | Polydipsia | 2 | 0.406 |
| C | Gender | 2 | 0.298 |
| C | Alopecia | 2 | 0.188 |
| C | visual blurring | 2 | 0.178 |
| C | Itching | 2 | 0.152 |
| C | partial paresis | 2 | 0.134 |
| C | Irritability | 2 | 0.120 |
| C | sudden weight loss | 2 | 0.086 |
| C | Polyphagia | 2 | 0.070 |
| C | Genital thrush | 2 | 0.064 |
| C | muscle stiffness | 2 | 0.042 |
| C | delayed healing | 2 | 0.036 |
| C | Obesity | 2 | 0.034 |
| N | Age | | 0.028 |
| C | weakness | 2 | 0.022 |

Fig.4 Likelihood Attributes with Corresponding Value.

We test the neural network based on the perception model, the model used for the neural network are the RELU, Tanh and Sigmoidal activation functions. Additionally, we included also the optimization solvers such as Adam, SGD, and L-BFGS-B for better accuracy results in an AUC, Accuracy, Precision, Recall, and Specificity. The following are the classification results of AUC, Accuracy, Precision, Recall, and Specificity. It shows that Adam and RELU got the highest AUC percentage of 99.3%. While Adam with Tanh provided the 98.9% AUC percentage, L-BFSG-B with Tanh got 98.3%, and SGD with RELU provided the minor AUC percentage with 92.1%. Additionally, the result for the accuracy, SGD with RELU got the highest percentage with 96.2%, and the second was the L-BFSG-B with Tanh having 95.7%.
The third highest accuracy was Adam with RELU with 95.5% compared to the least accuracy percentage got 89.7% with SGD and Sigmoid. Regarding the highest percentage for precision, L-BFSG-B with ReLu provides 96.2%. The second highest percentage was L-BFSG-B, with the having 95.7%. However, the most miniature precision is the SGD and Logistic having an 89.7%. Furthermore, regarding the recall, the highest percentage having a 96.2%, is L-BFSG-B and RELU. In comparison, the second-highest is L-BFSG-B, with Tanh

providing 95.7%. Moreover, a minor recall got 89.7% in SGD and Sigmoid. To sum up the result, Adam and RELU got the highest AUC percentage of 99.3%. With regards to accuracy, L-BFGS-B and ReLu provide 96.2%. Furthermore, L-BFGS-B and RELU got the highest percentage for precision with 96.2%. Moreover, L-BFGS-B and RELU also have the highest recall percentage providing 96.2%. Lastly, L-BFGS-B and RELU got the highest specificity percentage with 95.3%. The result for performance statistics was the majority to L-BFGS-B optimization solver and RELU activation function. The performance of the Neural Network for the proportion predicted is shown in Table 3.

Table 3: The Prediction Performance of ANN

| | | **Predicted (%)** | | |
|---|---|---|---|---|
| | | **Positive** | **Negative** | **Samples** |
| **Actual (%)** | **Positive** | 3.1 | 96 | 289 |
| | **Negative** | 96.9 | 4 | 179 |
| | **Samples** | 291 | 177 | 468 |

## 4. CONCLUSIONS

The best accuracy classification result is essential for the early stage of diabetes. It will prevent the patient from severe diabetic damage that causes damage to many parts of the body, including the eyes, heart, feet, nerves, and kidneys. In the worst scenario, it can cause fatality. However, experts can prevent this situation. The advent of new technology aligned with machine learning applications can help improve the early detection of diabetes.

In this study employed an artificial neural network (ANN) classification model. It has been tested with the Early-Stage Diabetes dataset. Each method has different parameters for accuracy, AUC, recall, precision, and specificity to determine the best classification results. In this study, the outstanding result that attained the highest accuracy percentage of 99.3% was the model with the RELU activation function for testing and training data and feature selection to optimize the accuracy result based on a Relief-Based Filter and obliterate the unlikelihood attribute. Additionally, the proposed study handled the overfitting and gave the best results for unbiased prediction using ROC analysis. Adding more instances and attributes will help the study improve its performance.

Additionally, this can help the medical experts provide proper treatment to the patient who suffers from the early stage of diabetes. This study demonstrates an experiment with the application of machine learning models in early-stage diabetes classification. Hence, it can help decrease the hazardous and prevalence of the early stage of diabetes.

## 5. REFERENCES

1. Roglic, G. (2016). WHO Global report on diabetes: A summary. International Journal of Noncommunicable Diseases, 1(1), 3.
2. Aishwarya, R., Gayathri, P., Jaisankar, N., 2013. A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology (IJET) 5, 2903–2908.
3. Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5–10.
4. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., 2017. Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal 15, 104–116. doi:10.1016/j.csbj.2016.12.005.
5. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.
6. Fatima, M., Pasha, M., (2017). Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications 09, 1–16. doi:10.4236/jilsa.2017.91001.
7. Kumar, P.S., Umatejaswi, V., (2017). Diagnosing Diabetes using Data Mining Techniques. International Journal of Scientific and Research Publications 7, 705–709.
8. Islam, M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques. In Computer Vision and Machine Intelligence in Medical Image Analysis (pp. 113-125). Springer, Singapore.
9. Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.
10. Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., ... & Abbas, Z. (2019). A model for early prediction of diabetes. Informatics in Medicine Unlocked, 16, 100204.
11. Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big data, 6(1), 1-19.
12. Kandhasamy, J. P., & Balamurali, S. J. P. C. S. (2015). Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 47, 45-51.
13. Gamara, R. P. C., Bandala, A. A., Loresco, P. J. M., & Vicerra, R. R. P. (2020, December). Early Stage Diabetes Likelihood Prediction using Artificial Neural Networks. In 2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM) (pp. 1-5). IEEE.
14. Le, T. T., Urbanowicz, R. J., Moore, J. H., & McKinney, B. A. (2019). Statistical inference Relief (STIR) feature selection. Bioinformatics, 35(8), 1358-1365.
15. Das, T. K. (2015, October). A customer classification prediction model based on machine learning techniques. In 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) (pp. 321-326). IEEE.