Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 256 (2025) 942–948

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANagement / HCist – International Conference on Health and Social Care Information Systems and Technologies 2024

Comparison of Neural Network Architectures for Diabetes Prediction

Nathan Antonio Guerreiro[a,c], Rui Nijo[a], João Paulo Teixeira[a,b] *[0000-0002-6679-5702]

[a]*Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Campus de Santa Apolónia, Bragança, 5300-253, Portugal*

[b]*Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal;*

[c]*Federal University of Technology - Paraná, Curitiba, 80230-901, Brazil*

Abstract

Diabetes represents a significant global health challenge, with millions of individuals affected and substantial impacts on healthcare systems. In this study, we compare two neural network architectures for diabetes prediction: the Feedforward Neural Network (FFNN) and the Cascade-Forward Backpropagation Neural Network (CFBPNN). Utilizing the Diabetes Prediction Dataset, comprising 100,000 samples, and after a balanced result, 17,000 samples were obtained. The networks are trained using the Levenberg-Marquardt and Resilient Backpropagation algorithms, and performance metrics, including precision, sensitivity, specificity, accuracy, F1-score, and computational time, are evaluated. Results indicate that the FFNN architecture paired with the Levenberg-Marquardt algorithm demonstrates superior diagnostic prediction accuracy with 91,10%. However, this comes at the cost of longer computational time compared to the CFBPNN.

Keywords: Neural Network Architectures, Diabetes Prediction, Machine Learning.

* Corresponding author. Tel.: +351 273 30 3129.
E-mail address: joaopt@ipb.pt

10.1016/j.procs.2025.02.198

1. Introduction

Diabetes is one of the chronic diseases that pose a significant challenge to healthcare systems globally. According to data from the World Health Organization (WHO), it is estimated that over 400 million people worldwide are currently living with diabetes, and this number is projected to increase in the coming years [1]. In Europe, there are approximately 60 million people with diabetes, which translates to around 10.3% of men and 9.6% of women aged 25 years and over [2]. Diabetes is also a public health issue, affecting millions of people and imposing a substantial burden on both patients and healthcare systems.

Given this concerning reality, the search for effective methods of diabetes prediction and management becomes increasingly urgent. Significant advances are being made in the healthcare field through the use of machine learning technologies, such as artificial neural networks (ANN). These techniques have been applied in various ways in medicine, ranging from disease prediction to the characterization of complex patterns in medical data.

The use of neural networks and machine learning in healthcare has shown promising potential to improve early diagnosis, personalized treatment, and disease prevention, including Alzheimer's disease [3], [4] and diabetes [5].

The authors in [6] applied Artificial Neural Networks (ANN) to identify Parkinson's, Huntington's, and Amyotrophic Lateral Sclerosis based on signals from the Electroencephalogram (EEG). They explored four types of methodologies but did not achieve the expected results due to the lack of a large dataset.

Another form of use of an ANN is in the form of time series prediction algorithms, as Oliveira et al [7] applied to predict seven days for cases and deaths of COVID-19 in Brazil, Portugal, and the United States. Several areas of health can use machine learning concepts with different methods. Guedes et al [8] show Long Short-Term Memory (LSTM) and ANN in the classification of chronic laryngitis, LSTM had a better performance with 100% accuracy, sensitivity, and specificity in the test set. Other areas of society can benefit from Machine Learning as tourism, the authors [9] investigate the predictive capacity of ANN methodology using datasets of international tourists visiting Vietnam from 2008 to 2020.

Through the analysis of large clinical datasets, neural networks can identify subtle patterns that may escape human detection, allowing for a more comprehensive understanding of patients' health conditions. In this context, the objective of this work is to conduct a comparative study between two neural network architectures for predicting samples with or without diabetes. Through this analysis, we seek to identify which neural network approach offers the best performance in diabetes prediction, thereby contributing to the development of more effective tools for disease prevention and management.

The present paper is structured as follows: The second section provides an overview of the dataset used, data preparation techniques, and details about the neural network architectures employed. The results section presents the outcomes of the study, including performance metrics such as precision, sensitivity, accuracy, and F1-score from the two architectures. Finally, the discussion and conclusion bring the findings,of our study.

2. Material and Methods

In this section, we present an overview of the materials utilized and the methodologies employed in our study to seek to identify which neural network approach offers the best performance in diabetes prediction.

2.1. Dataset

The dataset utilized, known as the Diabetes Prediction Dataset [10], has been compiled and maintained by various healthcare providers. Its primary objective is to classify the presence or absence of diabetes. To ensure data integrity, extensive cleaning and preprocessing were conducted, eliminating incomplete and extraneous information. Comprising 100,000 samples and garnering over 50,000 downloads, the dataset consists of nine columns. This comprehensive dataset forms the cornerstone of our analysis, enabling us to explore predictive models for diabetes with robustness and reliability.

It contains demographic information on patients, with 59% female, 41% male, and a small "other" category comprising 18 samples. The age range spans from 0 to 80 years, although diabetes is typically diagnosed more frequently in older adults. Hypertension and heart disease are indicated with values of 0 or 1, where 0 signifies the

absence of the condition, and 1 indicates its presence. Smoking history is categorized as "No info," "Never," "Former," "Not Current," "Current," and "Ever," with 36% of samples lacking information. Body mass index (BMI) is included, with higher values correlating with an increased risk of diabetes, ranging from 10.16 to 71.55 in the dataset. HbA1c (Hemoglobin A1c) levels, measuring a person's average blood sugar over the past 2-3 months, are also provided. Elevated levels of HbA1c indicate a higher risk of diabetes, with values typically exceeding 6.5%. Blood glucose levels, reflecting the concentration of glucose in the bloodstream at a given moment, serve as a crucial indicator of diabetes risk when elevated. The dependent variable or output of the model is the presence (1), or absence of diabetes (0).

2.2. Data Preparation

A methodology was adopted that consisted of assigning discrete numerical values to represent both gender and smoking history. The male gender was coded as 0, while the female gender was coded as 1, samples without gender weren't used in this study. As for smoking history, a progressive scale was used to categorize various situations, based on their historical context. The ordinal encoding method systematically assigned integer values to each category, reflecting the temporal progression and intensity of smoking habits. The scale began with "No info" as the baseline, designated with the value 0, indicating a lack of information. Subsequently, categories advanced through "Never"(1), "Former"(2), "Not Current"(3), "Current"(4), and finally, "Ever"(5) as the highest value category. This structured approach not only categorizes smoking behaviors but also preserves the inherent order and progression inherent in smoking history assessments.

This approach was adopted to facilitate the statistical analysis of the data and enable the application of modeling and inference techniques to investigate the variables in question.

The dataset was then sorted so that individuals with diabetes were placed at the beginning of the dataset, totaling 8,500 patients diagnosed with diabetes without considering diabetes types. Thus, the dataset was balanced to have the same numbers of pathological and non-pathological individuals. Therefore, the first 8,500 control subjects were selected, resulting in a new dataset with 17,000 samples.

After balancing the data, the first 8 columns were selected to create the input matrix , and the last column was removed to create the target/output matrix. It was necessary to convert the matrices to the double variable type.

2.3. Neural Network Architectures

The neural network training began with two architectures, maintaining the activation functions for each neural network layer but changing the training functions. This work explored a neural feedforward network and a neural cascade-forward backpropagation network.

2.3.1. Feedforward Neural Network (FFNN)

The FeedForward Network Neural (FFNN), also known as a MultiLayer Perceptron (MLP), is distinguished by the direction of information flow among its layers [11]. The simplicity and direct flow are seen in Figure 1, which has only one flow of information, from the input layer to the output layer, passing through hidden layers.

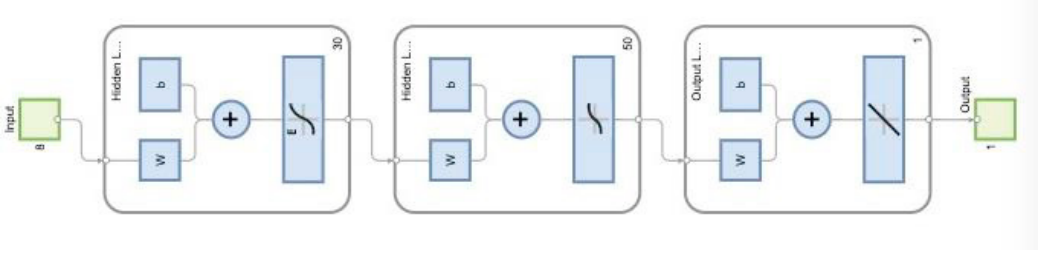

Fig. 1. Example of FFNN Architecture. Source: Authors' elaboration

2.3.2. Cascade-Forward Backpropagation Neural Network (CFBPNN)

In a Cascade-Forward BackPropagation Neural Network (CFBPNN), is similar to an FFNN in the use of the backpropagation algorithm. However, each layer of neurons is connected to all previous layers [12], as Figure 2 shows. This type of architecture gives the ability to have a memory of previous iterations [8].

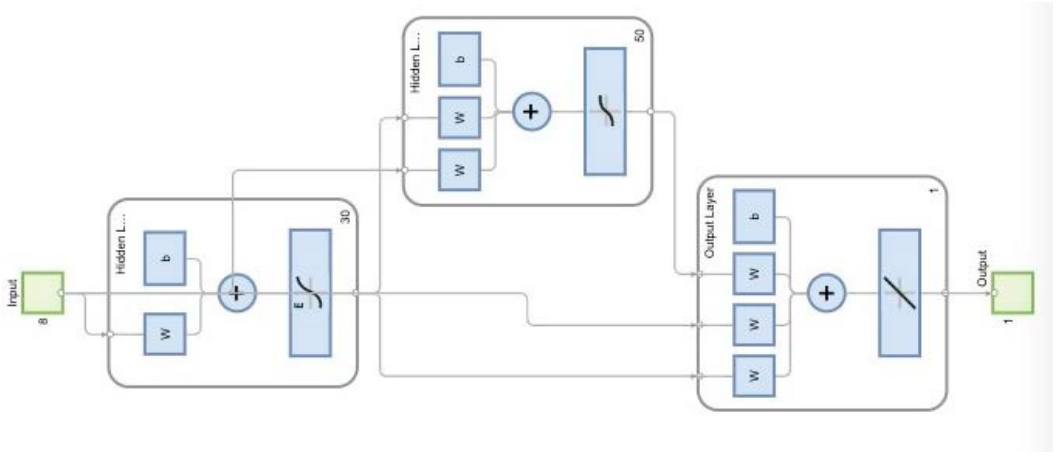

Fig. 2. Example of CFBPNN Architecture. Source: Authors' elaboration

Two hidden layers with 30 and 50 neurons, respectively, were used for both networks. The same activation functions for each layer of the neural network were maintained. In this case, the Elliot symmetric sigmoid transfer activation function was used for the first hidden layer, the Hyperbolic tangent sigmoid for the second hidden layer, and the Linear transfer function for the output layer. Figure 3 presents the activation functions [13 - 15].

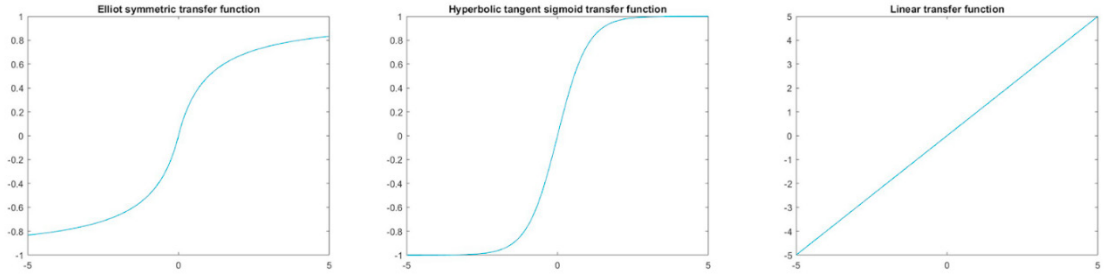

Fig. 3. Activations Transfer functions. Source: Authors' elaboration

2.4. Training and Evaluation

The neural networks were then trained using the prepared data. The dataset was divided into the default proportions, allocating 70% for training, 15% for validation, and the remaining 15% for testing.

Initially, the neural feedforward network was trained using the Levenberg-Marquardt training function, and after that, the Resilient backpropagation function was utilized. Performance metrics, including precision, sensitivity, specificity, accuracy, F1-score, and Elapsed Time, were collected, aiming for a high accuracy rate as it helps in better disease identification [16]. The same procedure was performed with the neural cascade-forward network, and the results were compared.

3. Results

After simulating the neural networks, the results were collected and organized as follows: Figure 4 shows the confusion matrix for the Levenberg-Marquardt training function, comparing FFNN on the right to CFBPNN on the left, where it is possible to see the True Positives, False Positives, False Negatives, and True Negatives.

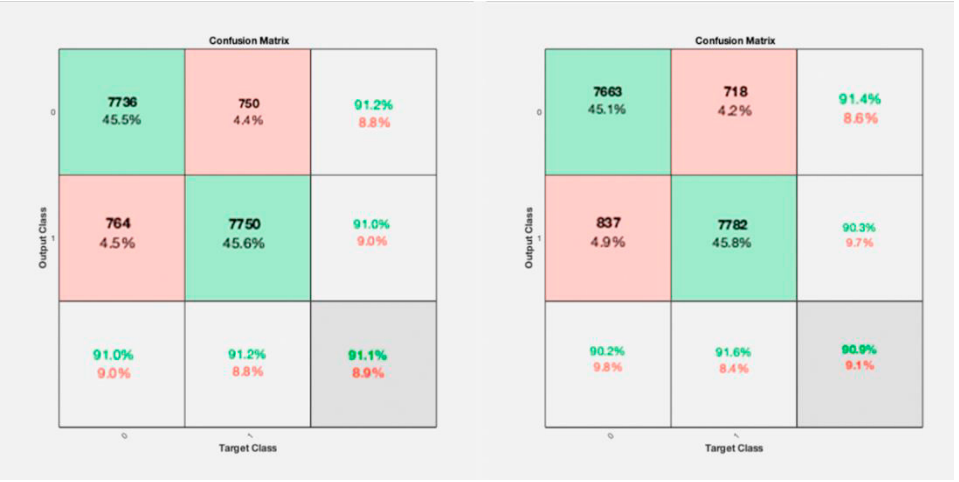

Fig. 4. Confusion Matrix for Levenberg-Marquardt train function. FFNN is on the right, and CFBPNN is on the left. Source: Authors' elaboration

For better analysis, Table 1 presents the results of the Levenberg-Marquardt training algorithm in percentage for precision, sensitivity, specificity, accuracy, F1-score, as well as Elapsed Time in seconds for training the two types of architectures.

Table 1. Comparison of mean evaluation metrics by Levenberg-Marquardt training algorithm.

	FFNN	CFBPNN
Precision	91,00%	90,20%
Sensitivity	91,20%	91,40%
Specificity	91,00%	90,30%
Accuracy	91,10%	90,90%
F1 Score	91,10%	90,80%
Elapsed time (s)	123	178

In the same way, the confusion matrix with the numbers of True Positives, False Positives, False Negatives, and True Negatives for the Resilient Backpropagation training algorithm was collected, as illustrated in Figure 5, comparing FFNN on the right to CFBPNN on the left.

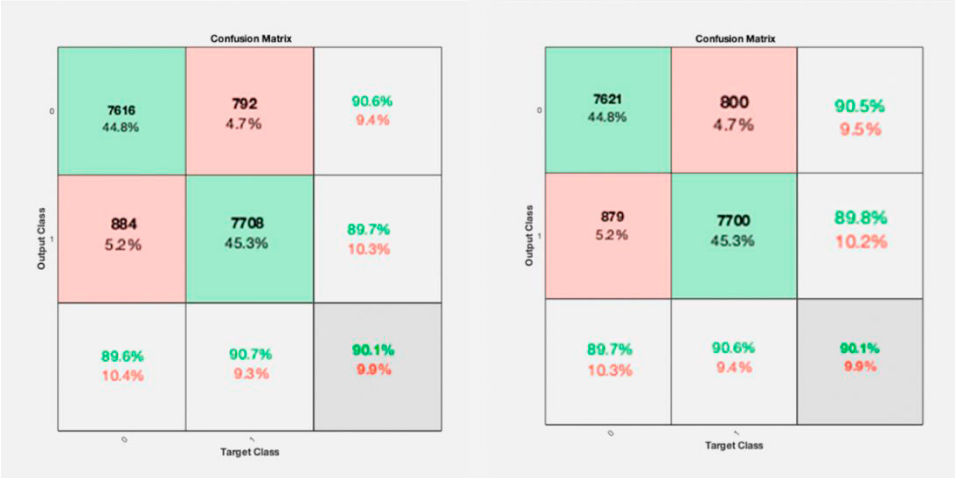

Fig. 5. Confusion Matrix for the Resilient backpropagation train function. FFNN is on the right, and CFBPNN is on the left. Source: Authors' elaboration

In percentage Table 2 presents precision, sensitivity, specificity, accuracy, F1-score, and Elapsed Time in seconds for training, all add for a better understanding of the Resilient backpropagation function.

Table 2. Comparison of architectures with evaluation metrics by Resilient backpropagation function.

	FFNN	CFBPNN
Precision	89,60%	89,70%
Sensitivity	90,60%	90,50%
Specificity	89,70%	89,80%
Accuracy	90,10%	90,10%
F1 Score	90,10%	90,10%
Elapsed time (s)	12	17

4. Discussion and Conclusions

In this study, we undertook a comparative analysis of two neural network architectures aimed at diabetes prediction. Both architectures were configured with identical numbers of hidden layers and neurons. The results gleaned from our simulations provide valuable insights. Specifically, our findings suggest that the Feedforward Neural Network (FFNN) architecture, when paired with the Levenberg-Marquardt training function, exhibits superior performance in diagnostic prediction, primarily due to its notably higher accuracy in comparison with the other architecture.

Noting that FFNN performed better in this case because there wasn't a temporal relationship between the data, which CFBPNN could have taken advantage of.

Although, it's worth noting that while the Levenberg-Marquardt training function yielded superior results, it also incurred a longer computational time compared to the Resilient backpropagation training function. This trade-off between accuracy and computational efficiency underscores the importance of carefully selecting the appropriate training function based on the application's specific requirements. Since the activation function of the output layer is linear, a limit value of 0.5 was considered, through experimentation, to consider the output as 0 (control) or 1 (diabetes). The values of metrics used, such as precision, sensitivity, specificity, accuracy, and F1-score, are very close to each other due to the dataset balancing. Therefore, there are no significant deviations in these metrics, as would occur if the dataset were not balanced. However, because we decided to balance the dataset, the results are slightly lower than those obtained using the entire dataset. The authors [17] using the same dataset, but maintaining the

imbalance between health and pathological cases, reported an F1-score of 98.5% evaluated on the complete dataset encompassing the training, validation, and test subsets. This approach differs from evaluating a balanced dataset. We would obtain higher accuracy at the expense of lower specificity and a difference between accuracy and F1-score.

Looking ahead, future studies could delve deeper into exploring alternative neural network architectures and training algorithms. By doing so, we may uncover novel approaches that not only enhance the accuracy of diabetes prediction but also optimize computational efficiency. This ongoing exploration is crucial for advancing the field of medical diagnostics and improving patient outcomes.

Acknowledgements

The authors are grateful to the Foundation for Science and Technology (FCT, Portugal) for financial support through national funds FCT/MCTES (PIDDAC) to CeDRI, UIDB/05757/2020 (DOI: 10.54499/UIDB/05757/2020) and UIDP/05757/2020 (DOI: 10.54499/UIDB/05757/2020) and SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020).

References

[1] Report of the WHO discussion group for people living with diabetes: virtual meeting, 30-31 March 2023. Geneva: World Health Organization; 2023. Licence: CC BY-NC-SA 3.0 IGO.

[2] World, “Diabetes EURO,” Who.int, Jan. 24, 2022. https://www.who.int/europe/health-topics/diabetes (accessed Apr. 10, 2024).

[3] Araújo, Teresa; Teixeira, João Paulo; Rodrigues, Pedro Miguel (2022). Smart-data-driven system for alzheimer disease detection through electroencephalographic signals. Bioengineering. ISSN 2306-5354. 9:4, p. 1-16. https://doi.org/10.3390/bioengineering9040141

[4] Pedro Rodrigues, João Paulo Teixeira; (2011) Artificial Neural Networks in the Discrimination of Alzheimer’s Disease; in M. M. C. Cunha, J. Varajão, P. Powell, R. Martinho (Editors) ENTERprise Information Systems Communications in Computer and Information Science, Volume 221, pp 272-281, Springer, ISBN 978-3-642-24351. DOI 10.1007/978-3-642-24352-3_29.

[5] Z. Soltani and A. Jafarian, “A New Artificial Neural Networks Approach for Diagnosing Diabetes Disease Type II,” International Journal of Advanced Computer Science and Applications, vol. 7, no. 6, 2016, doi: https://doi.org/10.14569/ijacsa.2016.070611.

[6] Rodrigues, P. M.; Teixeira, João Paulo. (2010) Classification of Electroencephalogram Signals Using Artificial Neural Networks. Proceedings of 3rd International Conference on BioMedical Engineering and Informatics (BMEI’10). Yantai – China, October 2010. DOI: 10.1109/BMEI.2010.5639941

[7 L. Sestrem de Oliveira, S. B. Gruetzmacher e J. P. Teixeira, “COVID-19 Time Series Prediction,” Procedia Computer Science, vol. 181, pp. 973–980, 2021. doi: 10.1016/j.procs.2021.01.254

[8] Guedes, V., Junior, A., Teixeira, F., Fernandes, J., & Teixeira, J. P. (2018) Long Short Term Memory on Chronic Laryngitis Classification, CENTERIS/ProjMAN/HCist 2018 - Procedia Computer Science – Elsevier, 138, 2018, Pages 250-257. https://doi.org/10.1016/j.procs.2018.10.036

[9] Nguyen, Le Q., Paula O. Fernandes, and João P. Teixeira. (2022). Analyzing and Forecasting Tourism Demand in Vietnam with Artificial Neural Networks Forecasting 4, no. 1: 36-50. https://doi.org/10.3390/forecast4010003

[10] M. Mustafa, “Diabetes prediction dataset,” Kaggle.com, 2023. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data (accessed Apr. 25, 2024).

[11] Zell, Andreas (1994). Simulation Neuronaler Netze [Simulation of Neural Networks] (in German) (1st ed.). Addison-Wesley. p. 73. ISBN 3-89319-554-8.

[12] Beale, M.H.; Hagan, M.T.; Demuth, H.B. Neural Network Toolbox User’s Guide; The MathWorks Inc.: Portola Valley, CA, USA, 1992; p. 103.

[13]“Elliot symmetric sigmoid transfer function - MATLAB elliotsig,” Mathworks.com, 2024. https://www.mathworks.com/help/deeplearning/ref/elliotsig.html (accessed Apr. 25, 2024).

[14]“Hyperbolic tangent sigmoid transfer function - MATLAB tansig,” Mathworks.com, 2024. https://www.mathworks.com/help/deeplearning/ref/tansig.html (accessed Apr. 25, 2024).

[15]“Linear transfer function - MATLAB purelin,” Mathworks.com, 2024. https://www.mathworks.com/help/deeplearning/ref/purelin.html (accessed Apr. 25, 2024).

[16] F.S. Gharehchopogh, Z.A. Khalifelu, "Neural Network Application in Diagnosis of Patient: A Case Study", IEEE, International Conference on Computer Networks and Information Technology (ICCNIT2011), Abbottabad, Pakistan, (2011), pp. 245-249.

[17] Dessanti, A., Pimentel, G. O., Teixeira, J. P (2024). A Comparative Analysis of MATLAB and Python Neural Networks for Diabetes Prediction. In Optimization, Learning Algorithms and Applications. OL2A 2024. Communications in Computer and Information Science. Springer.