LogNNet Neural Network Application for Diabetes Mellitus Diagnosis

Y. A. Izotov[1*], *M. T.* Huyut[2], and *A. A.* Velichko[1]

[1] Institute of Physics and Technology, Petrozavodsk State University, 185910 Petrozavodsk, Russia
[2] Department of Biostatistics and Medical Informatics, Faculty of Medicine, Erzincan Binali Yıldırım University, Erzincan 24000, Turkey

Abstract. The paper presents a LogNNet neural network algorithm for diabetes mellitus diagnosing based on a public dataset. The study used 100 thousand records of patient conditions. Model quality was evaluated using the Matthews Correlation Coefficient metric (MCC). The LogNNet neural network model showed high accuracy (MCC=0.733) in diabetes mellitus recognition. A highly positive relationship between HbA1c level and glucose level in the disease diagnosing was found using the LogNNet model. It has been observed that evaluating these variables together is much more effective than their individual effects in diagnosing the disease.

1 Introduction

Neural networks are a universal tool for solving tasks related to processing large volumes of diverse and incomplete diagnostic information in areas such as forecasting, modelling, control, optimization, and data analysis [1].

The relevance of reservoir computing (RC) research is determined by its effectiveness in solving forecasting tasks (weather changes, financial data) [2], equipment diagnostics and failure prediction [3], and nonlinear system control (robotics, automobiles, aircraft) [4].

The basic idea of RC is using a recurrent neural network acting as a reservoir with rich dynamics and powerful computational capabilities [5]. The weights of the reservoir are randomly generated, which eliminates the need for training [6]. Only the output neural networks (linear classifiers) are subjected to training, which are connected to the reservoir neurons through weight matrices.

RC has been successfully adapted to solving a lot of practical tasks related to real-world data. The simplicity of the training method in reservoir networks attracts researchers from related scientific fields. Most of these studies are related to traditional applications of machine learning methods: pattern recognition [7], system approximation [8], adaptive data filtering [9]. To improve the classification accuracy and enhance the approximation capabilities of reservoir networks, some studies employ a combination of multiple reservoirs, which increases the computational resource requirements.

A. Velichko has created a new neural network architecture, which is called LogNNet [10]. Its main feature is using deterministic chaotic filters for incoming signals, it means the system

* Corresponding author: izotov93@yandex.ru

tries to chaotically mix the input information, but at the same time extracts valuable data from the information that was initially invisible. A similar mechanism is used by reservoir neural networks. We used this network to classify the diabetes mellitus prediction dataset [11].

Diabetes mellitus (DM) is among the most common diseases worldwide [12]. Although it is not currently possible to cure the disease, attempts to control it are continuing.

Patients with DM may face other health problems, such as cardiac arrest or organ damage [12]. Therefore, early detection and management of DM will also prevent complications and help reduce the threat of serious health problems [13].

DM symptoms of in the initial stages are hard to identify. A result of advances in Artificial Intelligence (AI), early-stage detection of DM disease by an automated program is considered more likely and effective than the manual method of DM identification [14].

In this study, it is aimed to minimize the risks of human errors by reducing the workload of medical practitioners with the LogNNet neural network model, whose application has been demonstrated for the diagnosis of DM.

2 Methods

2.1 Diabetes mellitus prediction dataset

Diabetes mellitus prediction dataset is a set of data containing information about patients who have been tested for diabetes. In total, the database contains information about 100,000 patients. The database is publicly available [11] (Table 1).

Table 1. Diabetes mellitus prediction dataset list of attributes.

N	Attribute	Description	Range
1.	gender	Gender refers to the biological sex of the individual.	male, female and other.
2.	age	Biological age of the individual.	0 – 80
3.	hypertension	Refers to the existence of the hypertension condition.	0 or 1
4.	heart_disease	Heart disease is a medical condition that increases risk of developing diabetes	0 or 1
5.	smoking_history	Smoking history is also considered a risk factor for diabetes and can exacerbate the complications associated with diabetes	not current, former, No Info, current and never
6.	bmi	BMI stands for Body Mass Index, is a measure used to assess body weight relative to height. Underweight: BMI less than 18.5 Normal weight: BMI between 18.5 and 24.9 Overweight: BMI between 25 and 29.9 Obesity: BMI between 30 and 34.9 (Class I), 35 and 39.9 (Class II), and 40 or higher (Class III)	10.16 – 71.55
7.	HbA1c_level	HbA1c (Hemoglobin A1c) is a laboratory test that measures the average amount of glucose (sugar) in a person's blood over a period of approximately 2 to 3 months. It is primarily used to assess and monitor long-term blood sugar control in	3.5 – 9

		individuals with diabetes. Mostly more than 6.5% of HbA1c Level indicates diabetes.	
8.	blood_glucose_level	Blood glucose level refers to the concentration of glucose, a type of sugar, in your blood. The optimal levels usually fall between 70 and 130 mg/dL before meals, and less than 180 mg/dL two hours after starting a meal.	80 – 300

The data was collected using various medical studies and includes various patient characteristics such as age, blood glucose levels etc. Each patient in the dataset is divided into two classes by diagnosis: diabetes positive (1) or diabetes negative (0) based on the results of the test. The features used in the dataset are shown in Table 1.

The mission of the classification task is prediction of the presence or absence of diabetes in new patients based on their selected features.

The original dataset was transformed before the calculations began. The "gender" attribute was converted from a string representation of gender to a numeric representation of 0 (male), 1 (female) and 2 (other). The “smoking_history” attribute was also presented in numerical form 0 (not current), 1 (former), 2 (No Info), 3 (current) and 4 (never).

The result (Column “diabetes”) is also presented in binary form (diabetes mellitus positive (1), diabetes mellitus negative (0)). The dataset is unbalanced: diabetes mellitus is negative – 91,500 values, diabetes mellitus is positive – 8,500 values.

2.2 Architecture of LogNNet

LogNNet is a neural network based on the technology of “reservoir computing with auto-generation of weighting coefficients”. The LogNNet diagram is shown in Figure 1.

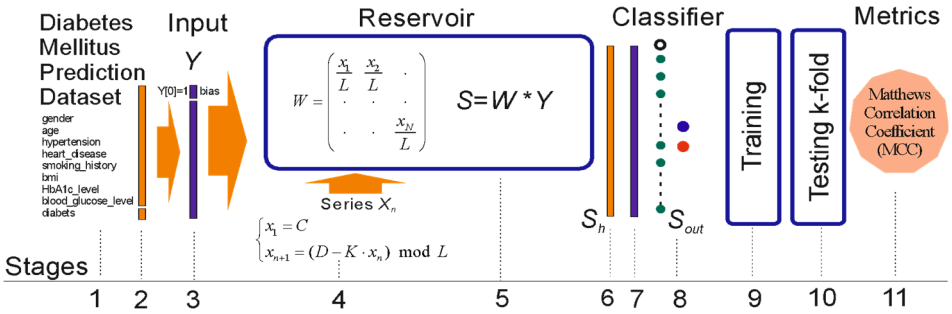

Fig. 1. LogNNet architecture and main stages of calculation.

The first step is reading data from the diabetes mellitus prediction dataset. Step 2 separates the resulting column (diabetes) into a separate array. Vectors Y from the dataset are sets of features from Table 1 with length = 8, followed by normalization (stage 3). Step 4 involves filling the reservoir matrix W using the congruent generator method. Next, the matrix W of dimension $(N+1)\oplus P$ is multiplied by the vector Y (stage 5). The result of the multiplication is a vector S_h of dimension P (stage 6), which is normalized and transformed into a vector S_h of dimension $(P+1)$ with a bias element $S_h[0]=1$ (stage 7). Vector S_h is passed into a two-layer linear classifier, while the hidden layer contains H neurons, and the output layer S_{out} contains $M=2$ neurons, which is equal to the number of possible classes (stage 8).

LogNNet is trained on a training data set with the number of epochs *Ep* (stage 9) and tested using the K-fold method (stage 10) described in section 2.3, the last step is calculating metrics (stage 11).

The particle swarm method was utilized for LogNNet hyperparameters selection. A total of 12 hyperparameters were involved. The first four hyperparameters L_1, L_2, L_3, L_4 corresponded to the coefficients of the linear congruent generator (step 4), responsible for the initial filling of the reservoir was carried out. The next five hyperparameters L_5, L_6, L_7, L_8, L_9 were used for optimizing the feature vector. Optimization included selection numbers of essential features that were extracted from the input vector and taken into account in further calculations. Features that were not taken into account in the calculation were eliminated by setting the coefficients of the reservoir matrix to zero. Hyperparameter L_{10} was responsible for the number of epochs of operation of the LogNNet neural network. Hyperparameter L_{11} stands for neurons number in the first hidden layer and hyperparameter L_{12} stands for neurons number in the second hidden layer of LogNNet neural network.

2.3 Classification metric

The metric used in this paper was Matthews Correlation Coefficient (MCC). It is a measure of the quality of classification predictions. It ranges from -1 to +1. A score of +1 indicates a perfect prediction, 0 indicates a random prediction, and -1 indicates a complete disagreement between the prediction and the actual outcome. The MCC is considered a robust measure because it takes into account imbalanced class distributions and is suitable for evaluating the performance of classifiers when the data is skewed.

3 Results and discussion

The distribution of the Pearson correlation coefficient between features is presented in Figure 2a. Figure 2b shows the correlation coefficient between diagnosis diabetes mellitus and features. The highest correlation of diagnosis is observed with attributes 7 and 8. In the right Figure, Figure 2b, the dependence of MCC on the feature number.

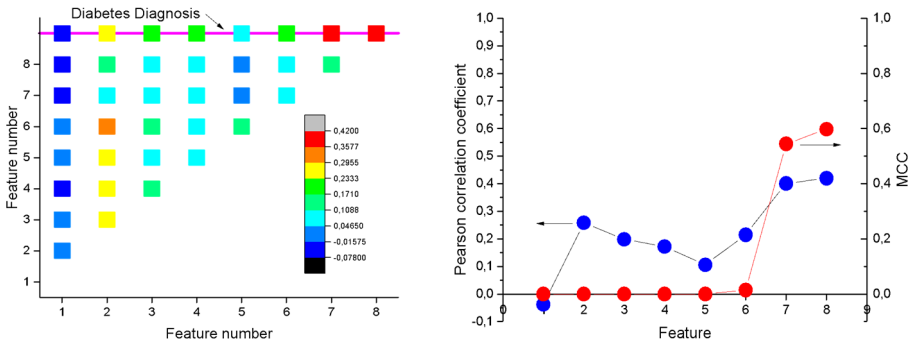

Fig. 2. Distribution of the Pearson correlation coefficient between features (a) and between diagnosis and features (b).

The results of the MCC assessment for individual attributes are presented in Table 2. It can be seen that the highest value corresponds to features 7 (HbA1c_level) and 8 (blood_glucose_level). Features 1-5 showed a zero MCC value, which indicates that the

diagnosis is independent of these features. Feature 6 (bmi) has a very minor MCC~0.015 effect on the diagnosis of diabetes.

The result of selecting the best combination of features is shown in Table 2. The combination of two features 7 and 8 showed the best MCC~0.733.

Table 2. Classification results of LogNNet neural network for single features.

Attribute	Matthews Correlation Coefficient	Best combination MCC=0.733
gender	0	0
age	0	0
hypertension	0	0
pheart_disease	0	0
smoking_history	0	0
bmi	0.014	0
HbA1c_level	0.544	1
blood_glucose_level	0.597	1

The results of comparing LogNNet neural network model and logistic regression in terms of Matthews correlation coefficient are presented in Table 3.

Table 3. Classification results of LogNNet neural network model and logistic regression model.

	LogNNet	Logistic regression
MCC	0.733	0.7055

The logistic regression model performed well. However, the LogNNet neural network showed slightly higher MCC value, which indicates a more accurate classification result when using this model.

One of the important results is that the dependence of the coefficient Pearson and MCC showed different results on secondary features 1-5, for example, for feature 2 (age) Pearson coefficient ~0.25 and MCC~0. Thus, a non-zero value of the Pearson coefficient does not mean that the feature is significant in the classification task, however, if the MCC has a high value, then with a high probability the Pearson coefficient will also be high, as shown by features 7 and 8.

4 Conclusion

In the past, the diagnosis of diabetes and measurement of elevated glucose levels typically involved fasting plasma glucose level (FPG), 2-hour plasma glucose (2HP), and random plasma glucose. Unfortunately, these methods could not provide information on long-term blood sugar levels. To tackle this issue, there has been a shift towards diagnosing diabetes and assessing its severity by measuring HbA1c and blood glycated proteins in the past decade. HbA1c is considered a crucial factor for diagnosing diabetes, although some studies suggest that it alone may not be enough to detect elevated glucose levels in diabetic patients.

In previous studies [15], HbA1c and blood glucose measurements have been stated to be a minimally invasive method to prevent diabetes complications. Additionally, diabetology and dentistry guidelines have reported that determining the relationship between HbA1c and glucose levels before implant treatment and oral surgery is an effective way to evaluate important complications [16]. As a matter of fact, it has been reported that poorly controlled DM patients are an important risk factor for many dental surgeries and various dental complications may occur [15]. However, many studies have reported that the role of the

relationship between HbA1c and glucose levels in the diagnosis of DM disease cannot be fully explained [17].

In this study, the LogNNet model was shown to diagnose DM disease with a high rate of accuracy by evaluating HbA1c and glucose levels together. In this context, LogNNet results seem promising for the earliest diagnosis of DM disease.

This research was supported by the Russian Science Foundation (grant no. 22-11-00055, https://rscf.ru/en/project/22-11-00055/, accessed on 30 March 2023).

References

1. I.H. Sarker, SN Comput. Sci. **2**, 154 (2021)
2. A.G. Salman, B. Kanigoro, Y. Heryadi, *Weather forecasting using deep learning techniques*, ICACSIS 2015 - 2015 International Conference on Advanced Computer Science and Information Systems, Proceedings (Institute of Electrical and Electronics Engineers Inc.), 281-285 (2016)
3. G. Tanaka, T. Yamane, J.B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, A. Hirose, Neural Networks **115**, 100-123 (2019)
4. X. Chen, T. Weng, C. Li, H. Yang, Phys. A Stat. Mech. its Appl. **607**, 128205 (2022)
5. H. Jaeger, GMD-Report 152, Ger. Natl. Res. Inst. Comput. Sci. (2002)
6. M. Lukoševičius, H. Jaeger, Comput. Sci. Rev. **3**, 127-149 (2009)
7. M. Liang, X. Hu, IEEE Computer Society Conference on Computer Vision and Pattern Recognition 3367-3375 (2015). https://doi.org/10.1109/CVPR.2015.7298958
8. S. Molavipour, H. Ghourchian, G. Bassi, M. Skoglund, Entropy **23**, 1-28 (2021)
9. L. Lu, C. Li, Z. Zhao, B. Bao, Q. Xu, Math. Probl. Eng. **2015** (2015)
10. A. Velichko, Electronics **9**, 1432 (2020)
11. Mustafa Mohammed, Diabetes prediction dataset (2023)
12. J. Chaki, Thillai Ganesh S, S.K. Cidham, Ananda Theertan S, J. King Saud Univ. - Comput. Inf. Sci. **34**, 3204-3225 (2022)
13. J. Ramsingh, V. Bhuvaneswari, J. King Saud Univ. - Comput. Inf. Sci. **33**, 1018-1029 (2021)
14. B. Sosale, S.R. Aravind, H. Murthy, S. Narayana, U. Sharma, S.G.V Gowda, M. Naveenam, BMJ Open Diabetes Res. Care **8**, e000892 (2020)
15. D. Végh, B. Bencze, D. Banyai, A. Vegh, N. Rózsa, Nagy Dobó C, Z. Biczo, G. Kammerhofer, M. Ujpal, Díaz Agurto L, I. Pedrinaci, Peña Cardelles J F, G.L. Magrin, N.M. Padhye, L. Mente, M. Payer, P. Hermann, Int. J. Environ. Res. Public Heal. **20**, 4745 (2023)
16. D. Busenlechner, R. Fürhauser, R. Haas, G. Watzek, G. Mailath, B. Pommer, J. Periodontal Implant Sci. **44**, 102-108 (2014)
17. T. Bomholt, B. Feldt-Rasmussen, R. Butt, R. Borg, M.H. Sarwary, T. Elung-Jensen, T. Almdal, F.K. Knop, K. Nørgaard, A.G. Ranjan, A. Larsson, M. Rix, M. Hornum, Nephron **146**, 146-152 (2022)