Diagnosis of Diabetes Mellitus Based on 11 Health Risk Factors using Backpropagation

F Arifin¹, R F Adhana², Shabrina Yois N A³, M Khairudin⁴ and N N M I Ma’arof⁵

¹,²,³,⁴ Departement of Electronic Engineering Education, Faculty of Engineering, Yogyakarta State University, Yogyakarta, Indonesia
⁵ Department of Engineering Technology, Faculty of Technical and Vocational, Universiti Pendidikan Sultan Idris, Malaysia

E-mail: ¹fatchul@uny.ac.id

Abstract. Diabetes Mellitus (DM) is a non-communicable disease and is a severe chronic illness. Early detection of DM is one way to detect the possibility of someone getting DM. This application was made to determine the accuracy of the diagnosis of DM using backpropagation ANN. There are 11 risk factors used, namely gender, smoker or not, heredity, systolic blood pressure, diastolic blood pressure, total cholesterol levels, HDL (High-Density Lipoprotein) levels, LDL (Low-Density Lipoprotein) levels, triglyceride levels, BMI (Body Mass Index), and HBA1c levels (Hemoglobin A1c). Risk factors are taken based on medical records of DM patients and data on healthy people. The training and testing of artificial neural networks showed promising results for the suitability of network output and desired targets with a correlation coefficient of 0.98043. The results of testing showed promising results for network output and target match desired with a correlation coefficient of 0.97894.

1. Introduction

Some dangerous diseases attack people. Diabetes is a dangerous disease characterized by increased levels of sugar in the blood that will flow to tissues throughout the body. Diabetes mellitus is a chronic hyperglycemic state and is slowly but surely going to destroy tissue within the body if not handled properly and seriously. Eventually, assorted chronic complications occur in the eyes, kidneys, nerves, and blood vessels [1].

Diabetes mellitus is a disease whose incidence rate is relatively high in many countries and is one disease that is a public health problem. People with diabetes mellitus in Indonesia in 2017 reached about 10.7 million and reaches more than 425 million people in the world. Of this amount is expected to continue to increase. Indonesia ranks sixth in the number of patients with diabetes mellitus in the world after India, China, and the United States [2].
Diabetes mellitus is closely related to a patient's blood sugar. For example, research conducted by Kudarti, et al, entitled "Early Detection of Diabetes Mellitus In Mothers PKK For High-Risk Pregnancy Prevention Efforts". Early detection of diabetes mellitus is done through the fasting blood sugar and blood sugar 2 hour post prandial which is complicated and time consuming [3].

In other side an increasing number of diabetes are often due to a delayed diagnosis of the disease. Therefore we need a low cost predictive tool that can provide fast diagnosis in determining whether a person has diabetes or not. Many methods are used to get the accurate prediction. This paper proposes a method of artificial neural networks that adopt the learning system in the human brain.

Several research of DM diagnostic using NN are has been done. Jefri Junifer Pangaribuan implement Extreme Learning Machine (ELM) with 8 pieces of input variables for diagnosis diabetes mellitus [4]. Another research is done by Rosita Sofiana applying optimization Backpropagation Adaptive Learning Rate and Momentum. Variables used as a data detection are physical symptoms and clinical experiences. There are 12 variables of symptoms used in this study [5].

In this research, we proposed backpropagation artificial neural network as our method. We use different variable for diagnostic which is expected to increase the accuracy of the prediction result. The backpropagation ANN based risk factors are implemented into software with a view Graphical User Interface (GUI). The development of this system to identify diabetes mellitus disease that can be used by the general public to check and find out what percentage of the likelihood of someone suffering from diabetes and providing an alternative convenience for the general public to conduct an initial assessment through the application.

2. Study Literature

2.1 Artificial Neural Networks

Artificial neural network (ANN) is one of the information processing system that is designed to mimic the way the human brain works in resolving a problem. ANN is able to identify activities based on past data. Past data will be studied by a neural network that has the capability to give a decision on the data that has never been studied [6].

Artificial neural network is one method of soft computing which inspired the development of the workings of human nerves. Implementation and research of artificial neural network running rapidly in various fields such as mapping pattern and pattern classification, image analysis and coding, signal processing, optimization, graphic manipulation, character recognition, robotics, processing knowledge and expert systems, health diagnosis, reduction of noise, etc [7]. One method that is used in the ANN is propagation.

Behind propagation neural network (Backpropagation) uses targeted training (supervised training) that facilitate in learning. By designing the training data in the form of pairs of input and target output pattern is desired, then the back propagation neural network (Backpropagation) will do the learning such that the result that comes closest to the training data.

The architecture of a network will determine the success of the targets to be achieved because not all problems can be solved with the same architecture. Network architecture is used by many layers (Multilayer network). Multilayer network architecture has one or more hidden layer that lies between the input layer and output layer (Figure 1). With the addition of one or more hidden layers, the ANN can solve more complex calculations. This can occur when the size of the input layer is quite large [4].

Training artificial neural networks are used, namely Supervised Learning. Mapping of input and output pairs are known. Input and output pair is used to train the network to obtain the desired weight. Couple these data serve as a "teacher" to train the network to obtain the best shape.

Methods of propagation or back propagation algorithm is one algorithm that is often used in solving the problem - the problem is complicated. This is because the network with this algorithm trained using supervised training. Backpropagation network training to get a balance between the network's ability to
recognize patterns used during training, and networking capabilities to provide the correct response to the input pattern similar (but not equal) to the pattern used during the training.

![Multilayer network diagram]

**Figure 1.** Multilayer network

### 2.2 Diabetes Mellitus

Diabetes mellitus is a disease characterized by high blood sugar levels caused by disturbances in insulin secretion or insulin or both disorders. The body of people with diabetes mellitus cannot produce insulin, which increases blood sugar levels. This can cause complications in these sufferers.

Diabetes mellitus (DM) is divided into several types. Type I diabetes usually causes no symptoms before the age of 30 years old patient, although symptoms can appear at any time. Type I diabetes patients require insulin from outside the body for survival. Type II diabetes usually experienced when patients aged 30 years or older, and patients are not dependent on insulin from outside the body, except in certain circumstances. Another DM mode is GDM, the DM that occurs in pregnant women, caused by impaired glucose tolerance in these patients. Currently, the number of patients with type II diabetes mellitus is increasing, because the pattern of increasingly unhealthy living, such as lack of physical activity and unhealthy eating patterns.

Epidemiological studies show that there are various conditions that accompany diabetes. The condition is called a risk factor for one or more of them is thought to increase a person's risk for experiencing DM, these several conditions including: (a) Gender, (b) Smokers / no, (c) Heredity factors, (d) Systolic blood pressure, (e) Diastolic blood pressure, (f) Total cholesterol levels, (g) HDL (High Density Lipoprotein), (h) LDL (Low Density Lipoprotein), (i) Triglyceride levels, (j) BMI (Body Mass Index), and (k) HBA1c levels (Hemoglobin A1c)

### 3. Methods

This study consists of several stages, the first study of literature is done to add references relating to the cases. Further observations include the analysis and retrieval of patient medical record patient data and data DM healthy people who check-up website [http://spandh.dcs.shef.ac.uk/gridcorpus/](http://spandh.dcs.shef.ac.uk/gridcorpus/). Once that is done the design and implementation using artificial neural network based on the risk factors of diabetes mellitus. Implementation is done in software and interfaces using the Graphical User Interface (GUI), then testing the software with test data that suffer from diabetes.
3.1. Input and Output Data Processing

Based on the data that has been obtained, there are 11 risk factors that serve as input variables, namely gender, smoker or not, heredity, systolic blood pressure, diastolic blood pressure, total cholesterol, HDL (High Density Lipoprotein), LDL (Low density Lipoprotein), high levels of triglycerides, BMI (Body Mass Index), and HBA1c levels (Hemoglobin A1c). These risk factors are used to diagnose how many percent chance of a person suffering from diabetes mellitus (DM). DM risk factor input variable used as a representation of a neural network in order to facilitate the reading of data. Tables 1, 2, and 3 show the classification and representation of gender, smoker or not and heredity based on the data obtained.

| Table 1. Classification and Representation of Gender |
|---------------------------------------------------|
| Variables | Representations |
| Gender   |               |
| Man      | 1             |
| Woman    | 0             |

| Table 2. Classification and Representation of Smoker or Not |
|----------------------------------------------------------|
| Variables | Representations |
| Smoker    |               |
| Yes       | 1             |
| Not       | 0             |

| Table 3. Classification and Representation of Factor Descendants |
|---------------------------------------------------------------|
| Heredity           | Variables | Representations |
| Factors            |           |                |
| There is           | 1         |
| Not                | 0         |

Tables 4 and 5 show the classification and representation of systolic and diastolic blood pressure (in mm Hg).

| Table 4. Classification and Representation of Systolic Blood Pressure |
|-----------------------------------------------------------------------|
| Systolic Blood | Classification of TDS | Representations |
| Pressure       |                       |                |
| <120           | Normal                 | 0              |
| 120-139        | Normal High            | 0.25           |
| 140-159        | Hypertension Low       | 0.5            |
| 160-179        | Hypertension High      | 0.75           |
| ≥180           | Hypertension Very high | 1              |

| Table 5. Classification And Representation of Diastolic Blood Pressure |
|---------------------------------------------------------------------|
| Pressure Diastolic blood | Classification of TDD | Representations |
|                         |                       |                |
| <80                     | Normal                 | 0              |
| 80-89                   | Normal High            | 0.25           |
| 90-99                   | Hypertension Low       | 0.5            |
| 100-109                 | Hypertension High      | 0.75           |
| ≥110                    | Hypertension Very high | 1              |
Tables 6, 7, 8, 9, 10, and 11 show the classification and representation of total cholesterol, HDL, LDL, triglyceride levels (in mg / dl), BMI, and HbA1C.

**Table 6.** Classification and Representation of Total Cholesterol Levels

| Total Cholesterol Levels | Classification of the summit | Representation |
|--------------------------|------------------------------|----------------|
| <5.2                     | Desired (still secure)       | 0              |
| 5.2 - 6.1                | High limit                  | 0.5            |
| ≥ 6.2                    | High                        | 1              |

**Table 7.** Classification and Representation of HDL Levels

| HDL Classification of HDL | Representations |
|---------------------------|-----------------|
| ≥1.0                      | Normal          | 0              |
| <1.0                      | Low             | 1              |

**Table 8.** Classification and Representation of LDL Levels

| Levels of LDL | Classification of LDL | Representation |
|---------------|-----------------------|----------------|
| <3.3          | Desired (still safe)  | 0              |
| 3.3 - 4.0     | High limit            | 0.3333         |
| 2.3 - 5.6     | High                  | 0.6667         |
| ≥5.6          | Very High             | 1              |

**Table 9.** Classification and Representation of Triglyceride Levels

| Triglyceride levels | Classification of Triglycerides | Representation |
|---------------------|--------------------------------|----------------|
| <1.7                | Normal                         | 0              |
| 1.7 to 2.2          | High limit                     | 0.3333         |
| 2.3 to 5.6          | High                           | 0.6667         |
| ≥ 5.6               | Very high                      | 1              |

**Table 10.** Classification and Representation of BMI (Body Mass Index)

| BMI values | Classification of BMI | Representations |
|------------|-----------------------|-----------------|
| <17.0      | Very thin             | 0               |
| 17.0 - 18.4| Thin                  | 0.25            |
| 18.5 - 25.0| Normal                | 0.5             |
| 25.1 - 27.0| Fat                   | 0.75            |
| > 27       | Very fat              | 1               |
Table 11. Classification and Representation of HbA1C

| HbA1C levels | Classification HbA1C | Representations |
|--------------|----------------------|-----------------|
| <5.7         | Normal               | 0               |
| 5.7 to 6.4   | Risk of Diabetes     | 0.5             |
| ≥ 6.5        | Indicated Diabetes   | 1               |

The variable output of the risk factors of diabetes mellitus (DM) is the percentage likelihood of someone suffering from diabetes from the range of 0% - 100%.

3.2. Artificial Neural Network Architecture

Back propagation neural network architecture that is used is a multilayer network. Multilayer network using one or more hidden layer located between the input layer and output layer. With the addition of one or more hidden layers, the neural network can solve more complex calculations. This can occur when the input layer size is large enough. Figure 2 shows the architecture (Multilayer network) based on the input layer, the first and second hidden layer, output layer.

```
- There are 11 inputs in the input layer:
  - X1 = Gender
  - X2 = smokers / no
  - X3 = Descent
  - X4 = (TDS) Systolic Blood Pressure
  - X5 = (TDD) Diastolic Blood Pressure
  - X6 = (of the summit) Total Cholesterol Levels
  - X7 = (HDL) cholesterol (High Density Lipoprotein)
```
• X8 = (LDL) cholesterol (Low Density Lipoprotein)
• X9 = (TG) levels of triglycerides
• X10 = (BMI) Body Mass Index
• X11 = HBA1c
- There is 1 hidden layer: Consists of 20 cells with 1 bias
- There is 1 neuron in Output layer

Table 12 is parameter configuration that is used for training and testing.

Table 12. Configuration JST That Used

| Parameter                          | Value         |
|-----------------------------------|---------------|
| The number of cells input layer   | 11            |
| Number of hidden layers           | 1             |
| The number of cells output layer  | 1             |
| constant learning                 | 0.09          |
| Target error                      | 0             |
| Momentum                           | 0.8           |
| Epoch                              | 100000        |
| activation function               | Logsig and purelin |

3.3 Designing the User Interface (UI)
Designing a user interface (UI) is of importance in the development of the computer application program. Therefore, the display of a computer program is a communication medium that will connect between users and applications. The program used to design a user interface (UI) using the GUI (Graphical User Interface) in software. Figure 3 shows DM diagnosis display of descriptions and exit the application. There are 11 entries available is gender, smoker or not, heredity, systolic blood pressure, diastolic blood pressure, total cholesterol, HDL (High Density Lipoprotein), LDL (Low Density Lipoprotein), high levels of triglycerides, BMI (Body Mass Index), and HBA1c levels (Hemoglobin A1c).

Figure 3. Display user interface of diagnosis DM
3.4 Training Neural Network

The training process in the capture of 48 training data that has been normalized based on the classification of sex, smoking, heredity, systolic blood pressure, diastolic blood pressure, total cholesterol, levels of HDL, the levels of LDL, triglyceride levels, BMI (Body Mass Index) and levels HBA1c will be used as input variables. With 3 seconds ANN generate 1166 training iterations with MSE 0.000111.

Equation for the best validation is 0.0016974 well as for a match between the output and target = 0.95 * T + 0.034 with a correlation coefficient (r1) is worth 0.98043 (approaching 1), has shown good results to match network output and the desired target. Can be seen in Figure 4 is the best picture performance validation of the ANN training and figure 5 is a relationship between the target and the output network to the training data.

![Figure 4. Best performance validation of training ANN](image)

![Figure 5. Relationship between target and output network for training](image)

Figure 6 output network ( ) and the target ( ) own or almost the same position. The best results are shown if the target and output in the same position exactly.
4. Results

4.1. Testing of Neural Network Sistem

The test data is done to determine whether the network is able to recognize the pattern of training data from a given input, if the error that has reached the target, then the resulting output can be used as the test data. New data testing performed on data that are not trained. The data used for the test data, there are 15 data is already represented. a diabetic patient the data and the data of healthy people who do a general check-ups are taken at random.

Equation to best fit = (0.95) T + (0.019) with a correlation coefficient of 0.97894 (approaching 1), has shown good results to match output to the target tissue (Figure 7).

Figure 6. Comparison between output target with the network for training data.

Figure 7. Relationship between target and output data network for testing

Figure 8 output network and the target almost at the same position. The best results are shown if the target and output the same position exactly. And the results of this test is said to be anywhere near accurate.
Figure 8. Comparing between the output target for data network testing

Training and testing are stored with siip.mat name that will be fed into the coding GUI as a reference so that applications can provide accurate answers.

4.2. Total Testing

Based on the risk factors for diabetes showed that percentage has been in testing into the application. Data tested the application into the test data is taken based on the classification of risk factors. The data will be tested into the application, as follows:
1. Gender Women, have a history of diabetes descent, a smoker, TDS 140 mmHg, 90 mmHg TDD, of the summit of 7.5 mmol / L, HDL 0.7 mmol / L, LDL 5 mmol / L, TG 1.1 mmol / L, 10.3% HbA1c and BMI 35. Testing Results in getting the percentage likelihood of someone suffering from diabetes amounted to 99.3769% as can be seen a woman has a risk factor value that exceeds normal limits.

Figure 9. Results first diagnosis

2. Gender - Men, do not have a history of diabetes descent, nonsmokers, TDS 110 mm Hg, 70 mm Hg TDD, of the summit of 3.2 mmol / L, HDL 2 mmol / L, LDL of 1.3 mmol / L, 0.3 mmol TG / L, HbA1c 4% and BMI of 17.2.
Testing Results in getting the percentage likelihood of someone suffering from diabetes amounted to 21.8854% due to the low value of the risk factors. Here is presented Figure 11 the results of testing the application on 63 patients consisting of patients are positive for diabetes and diabetic patients who were negative.

Based on the test results on 63 patients with positive and negative diabetes. The software application successfully classified the patients with positive DM and the patients with negative DM by prediction results value of above 50% and below 50% respectively.

5. Discussion
The developed system has some differences with the research and systems that have been made on previously works. However each research and systems have their own advantages and disadvantages. For example, in the systems made by [8], the study used architectural backpropagation (Backpropogation) uses 8 input layer and output layer with 2 hidden layers 5. Inputs used in the form of the number of times pregnant, 2-hour plasma glucose concentration in the oral glucose test, diastolic blood pressure, triceps skin fold thickness, 2 hours of serum insulin, BMI, heredity and age. With the output of the numbers 0 and 1 which if output is 0 then the patient is not suffering from diabetes whereas if output 1 patient suffering from diabetes. The results of this study indicate whether a patient is suffering from diabetes or not.

Another study conducted by [4] titled "Diagnosing Diabetes Mellitus Method Using Extreme Learning Machine". In that research, implement a new method of artificial neural network that is
Extreme Learning Machine (ELM). ELM is a neural network feed-forward with one hidden layer or better known as a single hidden layer feed-forward neural. This study uses the 8 pieces of input variables, including age, body mass index, blood pressure is lowest at any blood circulation, blood glucose concentration in the oral glucose tolerance test, number of pregnancies, diabetes asalasul function, skin fold thickness and serum insulin. This study uses the sigmoid activation function tan. ELM own predictions forecasting produce stable output with a range of the number of hidden neurons from 0 to 30 [6]. However, if the output produced ELM less than optimal, then the number of its hidden neurons will be changed.

The study in [5] has made "Diabetes Mellitus Early Detection System Using Neural Networks Backpropagation With Adaptive Learning Rate Optimization And Momentum". This study using Artificial Neural Network (ANN) by applying optimization Backpropagation Adaptive Learning Rate and Momentum. Variables used as a data detection are physical symptoms and clinical experiences. There are 12 variables of symptoms used in this study. The results showed that the best architecture for the detection Backpropagation obtained when using 1 hidden layer neurons with 9, 1000 epoch, and the error value targets 0.0001.

When compared with the three studies conducted previously by other authors, no one uses 11 risk factors DM as input cover gender, smoker or not, heredity, systolic blood pressure, diastolic blood pressure, total cholesterol, HDL (High Density Lipoprotein), LDL (Low Density Lipoprotein), high levels of triglycerides, BMI (Body Mass Index), and levels of HBA1c (Hemoglobin A1c) as the author made. In research conducted by Rosita Sofiana mentioned that used 12 input variables, but the source is not mentioned consist of any of the 12 variables. In contrast again with the system and study made by [4] and [8] that only uses 8 input variables.

In the neural network predictive systems that have been made include the output of a percentage of a person’s chances of DM disease indicated that the numbers displayed gives an overview of the percentage of the risk. In research and systems made by other researchers before, no one has made an application system diagnostics with an output in the form of a percentage. Systems that have been made previously by other authors, the output is generated only in the form of 0 and 1 is absolute so as not providing shadow indicated the percentage of a person’s risk of DM.

This diagnosis application system using reverse propagation method (propagation), similar to those that have been made previously by [8]. Although the method of propagation is considered good enough, but in a study conducted by [4] explained that speed prediction method Extreme Learning Machine (ELM) in the training data is 3.102 times faster than backpropagation and the testing of data 5136 times faster than backpropagation

6. Conclusion

Based on the results of research and discussion, it can be concluded as follows: First, the results of the training and testing the neural network of the 48 data show good results to match network output and the desired target with a correlation coefficient of 0.98043 approaches 1. And the new data test results as many as 15 data show good results for the output matching network and the target desired with a correlation coefficient of 0.97894 approaching 1. Second, Networks Neural Networks are able to diagnose diabetes mellitus (DM) is well based training and testing as many as 48 data, the testing of the data to the new data as many as 15 data, so the software can be used by users to diagnose what percentage of likelihood that a patient has the risk of diabetes mellitus

7. References
[1] A. D. Association, “Diagnosis and Classification of Diabetes Mellitus,” Diabetes Care, vol. 32, no. Suppl 1, p. S62, Jan. 2009, doi: 10.2337/dc09-S062.
[2] “Angka Diabetes di Indonesia Semakin Tinggi Berikut Faktanya.”
https://www.siloamhospitals.com:443/id/Contents/News-Events/Advertorial/2020/01/31/02/10/Angka Diabetes di Indonesia Semakin Tinggi Berikut Faktanya (accessed Sep. 22, 2020).

[3] Kudarti, et al, “Early Detection of Diabetes Mellitus In Mothers PKK For High-Risk Pregnancy Prevention Efforts,” *J. Holy Prodi DIII Midwifery Midwifery Acad. Mardi Rahayu*.

[4] J. J. Pangaribuan and Suharjito, “Diagnosis of diabetes mellitus using extreme learning machine,” in 2014 *International Conference on Information Technology Systems and Innovation (ICITSI)*, Nov. 2014, pp. 33–38, doi: 10.1109/ICITSI.2014.7048234.

[5] R. Sofiana and S. Sutikno, “Optimization of Backpropagation for Early Detection of Diabetes Mellitus,” *Int. J. Electr. Comput. Eng. IJECE*, vol. 8, no. 5, Art. no. 5, Oct. 2018, doi: 10.11591/ijece.v8i5.pp3232-3237.

[6] X. He and S. Xu, *Process Neural Networks: Theory and Applications*. Berlin Heidelberg: Springer-Verlag, 2010.

[7] M. Puri, A. Solanki, T. Padawer, S. M. Tipparaju, W. A. Moreno, and Y. Pathak, “Chapter 1 - Introduction to Artificial Neural Network (ANN) as a Predictive Tool for Drug Design, Discovery, Delivery, and Disposition: Basic Concepts and Modeling,” in *Artificial Neural Network for Drug Design, Delivery and Disposition*, M. Puri, Y. Pathak, V. K. Sutariya, S. Tipparaju, and W. Moreno, Eds. Boston: Academic Press, 2016, pp. 3–13.

[8] Sriyanto S and Sutedi S, Identifikasi Penyakit Diabetes Millitus Menggunakan Jaringan Syaraf Tiruan Dengan Metode Perambatan-Balik (Backpropagation), *Jurnal Informatika* 10, 2014, pp 79–94