Rangsit Journal of Arts and Sciences, January-June 2013 RJAS Vol.3 No. 1, pp. 53-63

ISSN 2229-063X (Print)/ISSN 2392-554X (Online)
DOI : 10.14456/rjas.2012.22

Application program for prediction of the type 2 diabetes in Thai people using artificial neural network

Wuttichai Luangruangrong[1*], Annupan Rodtook[2], and Sanon Chimmanee[3]

[1]IT Department, Bangkok Hospital, Bangkok 10310, Thailand
E-mail: wuttichai1053@hotmail.com
[2]Department of Computer Sciences, Ramkhamhaeng University, Bangkok 10240, Thailand
E-mail: sittisak168@yahoo.com
[3]Faculty of Information Technology, Rangsit University, Patumthani 12000, Thailand
E-mail: schimmanee@yahoo.com

*Corresponding author

Submitted 5 August 2012; accepted in final form 3 May 2013

Abstract

Among non-communicable diseases, diabetes kills the most people in Asia and is only becoming more prevalent in this region. Analyzing Type 2 diabetes risk factors utilizing prediction tools instead of blood testing is a challenge for accurate diabetes diagnosis. Recently, many researchers have studied the risk factors of diabetes by using Logistic Regression, Radial Basis and Back-propagation Neural Network (BNN) and applying them as a tool for diabetes prediction. This paper presents the new factors of smoking and alcohol consumption to improve performance in diabetes prediction. The predictive role of some traditional factors, i.e., body mass index, blood pressure and waist circumference and Family History (FMH) are also improved by adjusting the previously accepted ranges. The newly proposed diabetes prediction method is based on BNN. The sample data consists of 2,000 Thai people presenting at Bangkok hospital, Thailand from 2010 to 2012. From these experiments, it was found that an appropriate number of hidden nodes was equal to 50 nodes. Each proposed factor, i.e., FMH, alcohol consumption factor, smoking factor, and WC gave a better accuracy (correct in prediction) compared with a baseline model. Their accuracies were 83.35%, 83.50%, 83.60% and 83.65%, respectively. Subsequently, the new risk factor model performance was increased by tuning the neural network parameter learning rate. Our previously proposed factors for tuning BNN parameters introduced a high accuracy compared with the baseline model up to 1.2%. In this paper, the new proposed factors model introduces a better performance in Root Mean Square Error (RMSE) than the baseline factors model up to 25.75%, which are trained by the same sample data (2000 cases). Finally, the new model is implemented to be the diabetes prediction tool based on PHP web application, which works in conjunction with Matlab for predicting calculation. The threshold value returned is used to make a decision of whether or not patients have diabetes.

Keywords**:** *diabetes, prediction tools, back-propagation, neural network, neural network tuning, risk factors*

บทคัดย่อ

โรคเบาหวานเป็นโรคที่มีจำนวนผู้ป่วยเสียชีวิตมากที่สุดในเอเชีย และมีแนวโน้มเพิ่มสูงขึ้น การวิเคราะห์ปัจจัยความเสี่ยงของโรคเบาหวานชนิดที่ 2 มีความสำคัญมากสำหรับการพัฒนาเครื่องมือเพื่อช่วยประเมินความเสี่ยงแทนการตรวจเลือด ซึ่งมีงานวิจัยจำนวนมากที่ ศึกษาเกี่ยวกับปัจจัยความเสี่ยงของการเกิดโรคเบาหวานสำหรับการพัฒนาเครื่องมือดังกล่าวด้วย วิธีการถดถอยโลจิสติก ฐานรัศมี และการแพร่กระ จายแบบย้อนกลับ งานวิจัยฉบับนี้จึงนำเสนอปัจจัยความเสี่ยงใหม่คือ การสูบบุหรี่ และ การบริโภคแอลกอฮอล์ เพื่อเพิ่มประสิทธิภาพในการประเมินความเสี่ยงของโรคเบาหวาน และปรับเปลี่ยนปัจจัยความเสี่ยงเดิม เช่นดัชนีมวลกาย ความดันโลหิตตัวบน ความยาวเส้นรอบเอว และประวัติทางครอบครัว ด้วย โครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ ซึ่งใช้ข้อมูลของกลุ่มประชากรไทยจากโรงพยาบาลกรุงเทพจำนวน 2,000 ราย ในช่วงปี 2010 ถึง 2012 ในการเรียนรู้ของโครงข่ายประสาทเทียม ซึ่งจากการทดลองแสดงให้เห็นว่า จำนวนโหนดชั้นซ่อนที่เหมาะสมคือ 50 โหนด และปัจจัยที่นำเสนอคือ ประวัติทางครอบครัว การบริโภคแอลกอฮอล์ การสูบบุหรี่ และความยาวเส้นรอบเอว ได้ส่งผลให้ค่าความถูกต้องเพิ่มขึ้นเป็น 83.35%, 83.50%, 83.60% และ 83.65% ตามลำดับ และรูปแบบการนำเข้าปัจจัยเสี่ยงรูปแบบใหม่ได้ถูกนำมาปรับปรุงประสิทธิภาพด้วยการเพิ่มค่าอัตราการเรียนรู้ใหม่ซึ่งมีค่าน้อยลง โดยทั้งปัจจัยที่นำเสนอและการปรับปรุงประสิทธิภาพโครงข่ายประสาทเทียม ในงานวิจัยก่อนหน้านี้ จึงสามารถทำให้ค่าความถูกต้องสูงขึ้นเมื่อเทียบกับรูปแบบการนำเข้าปัจจัยที่ใช้เป็นฐานการวัดผล 1.2 % และสำหรับการนำข้อมูลเข้าทั้งหมดในการเรียนรู้ของโครงข่ายประสาทเทียม รูปแบบการปรับปรุงโครงสร้างการเรียนรู้ของโครงข่ายประสาทเทียมที่นำเสนอในงานวิจัยฉบับนี้มีประสิทธิภาพเพิ่มขึ้น 25.75% รูปแบบที่มีประสิทธิภาพสูงที่สุด ได้ถูกนำมาเพื่อพัฒนาเป็นเครื่องมือสำหรับพยากรณ์โรคเบาหวานในรูปแบบโปรแกรมประยุกต์บนเว็บไซต์ด้วยภาษาพีเอชพี โดยประมวลผลควบคู่กับโปรแกรมแมตแล็บซึ่งจะทำหน้าที่คำนวนค่าในการพยากรณ์

คำสำคัญ: *โรคเบาหวาน, เครื่องมือพยากรณ์, การแพร่กระจายย้อนกลับ, โครงข่ายประสาทเทียม, ปัจจัยความเสี่ยง*

1. Introduction

Diabetes is increasing in pandemic proportions even more so than HIV/AIDS. There are two types of diabetes. Type 1 diabetes is an unpreventable autoimmune disorder, where the body destroys the insulin-producing beta cells in the pancreas. However, Type 2 diabetes tends to appear later in life and may depend on controllable risk factors (Centers for Disease Control and Prevention, 2007). Many researchers have recently studied the risk factors of diabetes in order to develop diabetes forecasting tools, with simple datasets based on Netherlands, Denmark, India and United States of America (Baan, Ruige, Stolk, Witteman, Dekker, Heine, & Feskens, 1999; Glümer, Carstensen, Sandbaek, Lauritzen, Jørgensen, & Borch-Johnsen, 2004; Stern, Williams, González-Villalpando, Hunt, & Haffner, 2004; Schmidt, Duncan, Bang, Pankow, Ballantyne, Golden, Folsom, & Chambless, 2005; Mohan, Deepa, Deepa, Somannavar, & Datta, 2005; Wilson, Meigs, Sullivan, Fox, Nathan, & D'Agostino, 2007). The studies from Aekphakorn (2005) and Wisaeng , Chiewchanwattana, and Sunat (2009) developed the simple dataset for the Thai population based on diabetes related risk factors. These researchers used a Logistic Regression, Radial Basis and Back-propagation Neural Network (BNN) by evaluating the following risk factors: age, gender, BMI, blood pressure (BP), Family History (FMH), race, waist circumference (WC) and physical activity.

Clair, Bitton, Meigs, James, and Rigotti (2011) found that smoking and blood glucose levels are related, which is illustrated by classifying the cotinine categories. Conversely, studies by Rimm, Manson, Stampfer, Colditz, Willett, Rosner, Hennekens, and Speizer (1993) and Rimm, Chan, Stampfer, Colditz, and Willett (1995) assessed cigarette smoking and alcohol consumption as the risk factors for diabetes in men and women. It was shown that cigarette smoking may be an independent factor, while both genders who consumed higher amounts of alcohol may have a reduced risk of diabetes. Importantly, many papers attempted to prove that alcohol consumption is an effective means to reduce the risk of diabetes. Koppes, Dekker, Hendriks, Bouter, and Heine (2005) posited that moderate alcohol consumption equates to a lower risk of diabetes. An important caveat is that heavy drinkers have a higher risk of diabetes, which is equal to non-drinkers. Baliunas, Taylor, Irving, Roerecke, Patra, Mohapatra, and Rehm (2009) found that the relative risk of alcohol consumption has a U-shaped relationship for both genders.

For the traditional risk factors, the Centre for Genetics Education (CGE) at Royal North Shore Hospital in Sydney proposed an effect of inherited predisposition of diabetes in brother, sister, parent, child and identical twins which differed possibly based on European data, which lead to split FMH variable into father, mother and relative history (Barlow-Stewart, 2007). Recently, there were two ranges of BMI proposed from the World Health Organization (2012) and by Asian American Diabetes Initiative (Joslin Diabetes Center, 2012) that differ from traditional BMI ranges. The BP factor also had a new range proposed by Golden, Wang, Klag, Meoni, and Brancati (2003). Additionally, the WC factor had new ranges from work by Matoba, Inoguchi, Nasu, Suzuki, Yanase, Nawata, and Takayanagi (2007) and Matsuzawa (2005).

Our research proposed a diabetes prediction method based on BNN, which used diabetes risk factors from newly updated BP, BMI and WC factors (Rimm et al., 1993; Rimm et al., 1995; Koppes, et al., 2005; Baliunas, Taylor, Irving, Roerecke, Patra, Mohapatra, & Rehm, 2009). In addition, new ranges for traditional risk factors (Barlow-Stewart, 2007; World Health Organization, 2012; Joslin Diabetes Center, 2012; Golden et al., 2003; Matoba et al., 2007; Matsuzawa, 2005) are evaluated with BNN method in this paper to improve their performance. To further improve accuracy tuning of BNN parameters, i.e., learning rate and epochs, is necessary. 2,000 patients with diabetes from Bangkok Hospital are used for training in order to find the appropriate factor and the range for developing the diabetes risk prediction model.

From results of three experiments two new factors for diagnosing diabetes in Thai people have been found: alcohol with U-shaped relationship and smoking. The proposed range adjustment of traditional factors also gave a higher performance. Additionally, the proposed tuning of BNN parameters introduced a lower Root Mean Square Error (RMSE). Therefore, our previously proposed model in (Luangruangrong, Rodtook & Chimmanee, 2012) gave a better performance compared with a baseline model (Wisaeng et al., 2009) up to 1.2%.

This paper intends to extend the previously proposed model to enhance performance up to

25.75% and to construct a web application for diabetes prediction. Additional experiments will combine the proposed model factors from experiment 2 and the learning rate parameters from experiment 3 (Luangruangrong et al., 2012) into a novel implemented model as shown in experiment 4. From these experimental results, RMSE values of the proposed model are shown to be better than the baseline model (Wisaeng et al., 2009) with the same proposed learning rate parameters in experiment 3, decreasing from 0.5617 to 0.4171, indicating that the performance is increased 25.75%. The novel PHP web application uses Matlab for predicting calculations.

The paper will be organized as follows: In section 2, we state related and previous works. Section 3 presents the proposed method. Section 4 gives the experimental results. Section 5 offers conclusions and future work.

2. Literature review

Traditionally, diabetes has two types. Type 1 diabetes mellitus is an organ-specific T-cell–mediated autoimmune disease characterized by cellular infiltration of pancreatic islets and destruction of insulin-producing beta cells. Type 2 diabetes tends to occurs later in life and its occurrence is based on several risk factors (Centers for Disease Control and Prevention, 2007). This section is divided into 2 parts: Related work and Previous work.

2.1 Related work

2.1.1 Prediction methods for diabetes diagnosis

Many researchers have studied and found risk factors for diabetes. Most used a method of the logistic regression (LR) (Aekphakorn, 2005; Baan et al., 1999; Glümer et al., 2003; Stern et al., 2004; Schmidt et al., 2004; Mohan et al., 2005; Wilson et al., 2007). Additional groups, Wisaeng et al. (2009) evaluated methods for the risk factors of developing diabetes that consisted of LR, radial basis and BNN.

Baan et al. (1999) developed a model with LR for identifying people with undiagnosed diabetes in a Netherlands sample data set of 1,016 people. Glümer et al. (2004) developed a questionnaire to predict the risk of diabetes prevalence in the population of Denmark, with a sample data set of 6,784 people. Stern et al. (2004) studied a population in San Antonio consisting 1,791 people of Mexican heritage and 1,112 Caucasians. Schmidt et al. (2005) developed and evaluated clinical rules to predict the risk of diabetes for 7,915 middle-aged Americans.

Mohan et al. (2005) developed and validated a simplified Indian diabetes risk score for detecting undiagnosed diabetes in 26,001 Indian people. Wilson et al. (2007) used the logistic regression model to predict incident diabetes by using a sample population which contained 3,140 men and women in America. This estimated the risk of new Type 2 diabetes occurrences during a 7-year follow-up interval.

Recently, Thai researchers also studied this area of diagnosis. Wisaeng et al. (2009) assessed the diabetes risk model for medical diagnosis in cases of Thai people by using BNN, Radial Basis and LR, which showed that BNN is the most accurate. Aekphakorn (2005) developed an assessment questionnaire for diabetes risk in Thai people (sample size of 2,677) who are working in the Electricity Generating Authority of Thailand by using LR. Table 1 lists a summary of risk factor and method of the researchers as mentioned above.

Table 1 A summary of the risk factors and methods for diabetes

Researcher	Factors	Analytical method
Baan, 1999	Model 1: Age, Sex, Obesity, BP	Logistic Regression
	Model 2: Add Family history(FMH) and BMI	
Stem, 2002	Age, Sex, BMI, BP, FMH, Race	
Glümer, 2004	Age, Sex, BMI, BP, FMH	
Wichai, 2005	Age, Sex, BMI, BP, WC, FMH	
Schmidt, 2005	Age, Height, BP, FMH, Race, WC	
Mohan, 2005	Age, FMH, WC, Physical Activity	
Wilson, 2007	Age, Sex, FMH, BMI	
Krittipon, 2009	Age, Sex, BMI, BP, FMH	Logistic Regression
		Radial Basis
	Age, Sex, BMI, BP, WC, FMH	BNN

2.1.2 Traditional diabetes risk factors

A blood test is the only method to conclusively diagnose diabetes. However, it is costly and not convenient. Instead, physicians attempt to predict diabetes by utilizing known risk factors, patient history and any clinical symptoms that are presented. The traditional risk factors for diabetes diagnosis are age, BMI, gender, BP, FMH, and WC (Aekphakorn, 2005; Baan et al., 1999; Glümer et al., 2004; Stern et al., 2004; Schmidt et al., 2005; Mohan et al., 2005; Wilson et al., 2007; Wisaeng et al., 2009). For Thai people, the traditional factors (Wisaeng et al., 2009) are listed in Table 2.

Table 2 The traditional diabetes risk factors and their range for Thai people (Wisaeng et al., 2009) are used as a baseline in the paper called as a Set A.

Factor	Range	Factor	Range
Age(Years)		**BP**	
x<=44	-1	x<120	-1
44<x<50	-1 to 1	120<=x<140	-1 to 1
x>=50	1	x>=140	1
BMI		**FMH**	
x<23	-1	No	-1
23<=x<=27.5	-1 to 1	Yes	1
x>27.5	1		
Gender		**WC**	
Female	-1	Male x<90 and Female x<80	-1
Male	1	Male x>=90 and Female x>=80	1

2.1.3 The new risk factors for diabetes diagnosis: Smoking and alcohol consumption

Clair et al. (2011) found that smoking and blood glucose levels are related by using cross-sectional data based on classifying cotinine categories. Cotinine is an important metabolite in the metabolism of nicotine. Using similar methodology, Rimm et al. (1993) and Rimm et al. (1995) studied the risk factor of cigarette smoking in men and women. It was found that cigarette smoking may be an independent factor. Additionally, Rimm et al. (1993) and Rimm et al. (1995) found that alcohol consumption is a diabetes risk factor for both genders. The most significant finding was that moderate alcohol consumption reduced the risk of diabetes.

Koppes et al. (2005) studied alcohol consumption as a diabetes risk factor by meta-analysis of prospective observational studies and indicated that moderate drinking (<48 g/day) lowers diabetes risk. Heavy drinkers (≥48 g/day) are equal to non-drinkers. Recently, Baliunas et al. (2009) found the risk associated with alcohol consumption is in a U-shaped relationship for both genders. Among men and women the daily alcohol consumption amounts were most protective when consuming 22 and 24 g/day, respectively and became deleterious at just over 60 and 50 g/day, respectively.

2.1.4 Range adjustment of the traditional factors

• BMI - World Health Organization (2012) suggested the cut-off point between 22 and 26. Additionally, Asian American Diabetes Initiative (Joslin Diabetes Center, 2012) suggested 18.5, 24 and 27.

• BP - Golden et al. (2003) found individuals who subsequently developed diabetes with baseline BP >130 mmHg, had a 25% increased risk of developing diabetes compared with those with baseline BP <130 mmHg.

• WC - Matoba et al. (2007) reviewed cross-sectional data from 1,658 men and 1,116 women from the results of annual medical checkup during May 2005 to November 2006 at the Human Dry Dock Center Wellness in Fukuoka, Japan. The checkups determined optimal waist cut-off points of 87 cm in men and 80 cm in women. The Japanese committee (Matsuzawa, 2005) decided to adjust waist circumference cut-off point as 85 cm for men and 90 cm for women, because amounts of subcutaneous fat are greater in women with the same visceral adiposity.

• FMH - Barlow-Stewart (2007) proposed estimated risks for developing diabetes based on European data. The impact of inherited predisposition is 10% for brother or sister, 10%

parent, 20% for brother or sister and a parent or child and 50% for identical twins.

2.2 Previous work

Our previous work is presented in IEEE SMC 2012, Korea (Luangruangrong et al., 2012), which proposed the new risk factors model and BNN tuning as mentioned in section 3. In this paper we extend our previous work by combining experiments 2 and 3 into experiment 4 that compares a value of RMSE between the new factors model and baseline model. The neural network parameters i.e., weights and bias are used in Matlab for the predicting calculation that is called by PHP web application as shown in Figure 1.

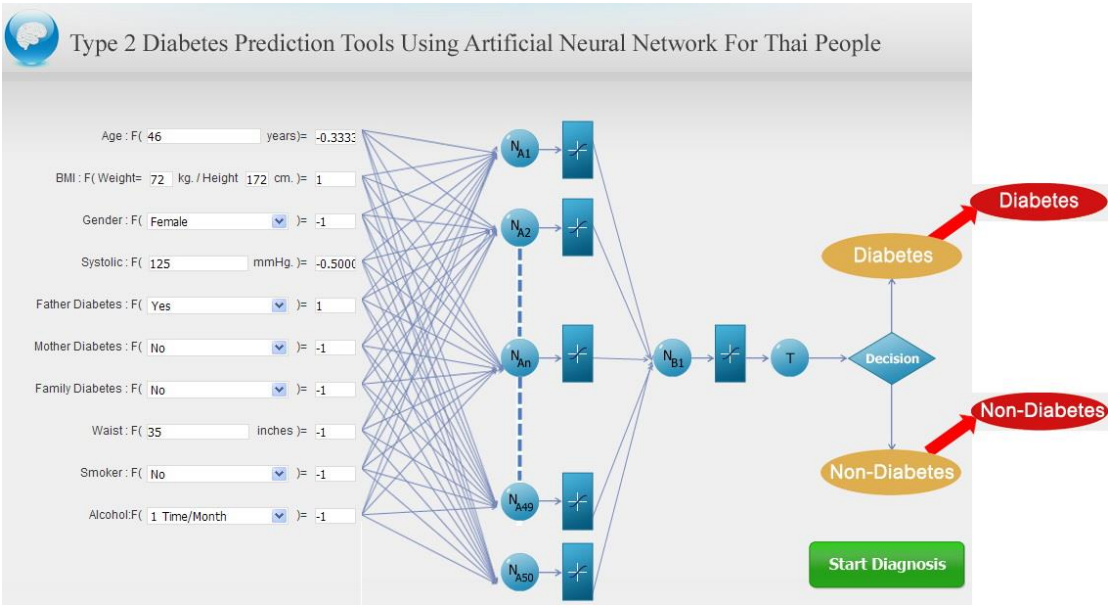

Figure 1 The web application user interface and the prediction presentation

3. The proposed method

This paper presents a novel method of the diabetes diagnosis based on BNN which is extending from previous work. The proposed method is divided into two parts. First, new additional factors are considered while the range of traditional factors is adjusted. Subsequently, tuning BNN parameters (learning rate, and epochs) are done. Subsection D is an extension of the previous work in Luangruangrong et al. (2012), which is developed from the outcome of experiment 4 as listed in Table 8.

3.1 The architecture of BNN for diabetes prediction

In Figure 2, there are three layers in BNN, consisting of input, hidden and output layers. The input layer is composed of 10 nodes, which is shown in subsection 4.b.2. The hidden layer contains 50 nodes (more detail can be found in subsection 4.2.1). For the learning state, weights are randomized initially in a range of -0.5 to 0.5. The output layer is a single node. All transfer functions are Tan-sigmoid. In this paper, the RMSE is used to find appropriate BNN structure and risk factors.

3.2 The proposed risk factors

This section consists of two parts. In part one, the new risk factors will be proposed. The second part will contain a range adjustment of traditional risk factors.

3.2.1 The proposed risk factors

This paper presents two new diabetes risk factors: smoking and alcohol consumption. Smoking factor is referenced in Clair et al. (2011) here known as set H. Alcohol factor has two methods therefore, there are two sets, which are called as set I (Rimm et al., 1993; Rimm et al., 1995) and set J (Koppes et al., 2005) in this paper. These new factors are listed in Table 3.

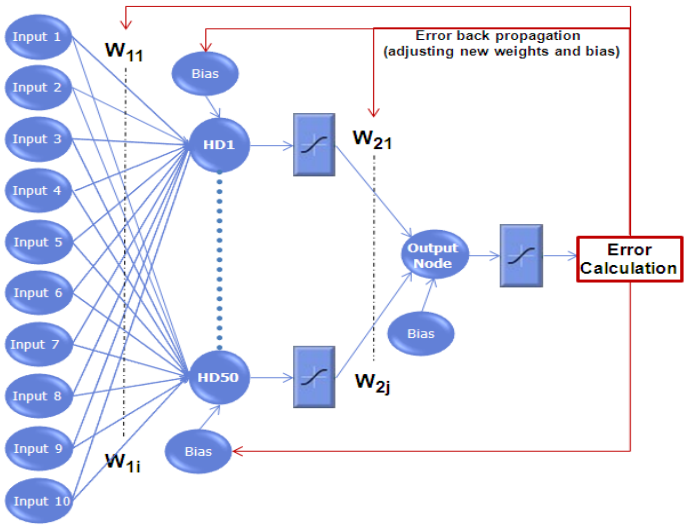

Figure 2 BNN architecture contains a hidden layer with 50 hidden nodes

Table 3 The new proposed factors for improving the prediction

Factor	Range	Factor	Range
Smoking (Set H)		**Alcohol (Set I)**	
Yes	-1	Yes	1
No	1	No	-1
		Alcohol/Week (Set J)	
		0<x<2	-1
		x=0 or 2<=x<=3	0
		x>=4	1

3.2.2 Adjusting the range of factors.

This paper also investigates the traditional factors which are BMI (World Health Organization, 2012; Joslin Diabetes Center, 2012), BP (Golden et al., 2003), FMH (Barlow-Stewart, 2007) and WC (Matoba et al., 2007; Matsuzawa, 2005). New ranges are proposed to increase the accuracy of diabetes prediction.

Table 4 The factor with new ranges of diabetes risk

Factor	Range	Factor	Range
BMI (Set B)		**WC (Set F)**	
x<18	-1	Male x<85 and Female x<90	-1
18<=x<24	-1 to 0	Male x>=85 and Female x>=90	1
24<=x<26	0 to 1		
x>=27	1		
BMI (Set C)		**WC (Set G)**	
x<22	-1	Male x<87 and Female x<80	-1
22<=x<26	-1 to 1	Male x>=87 and Female x>=80	1
x>=26	1		
BP (Set D)		**FMH (Set E)**	
x<130	-1	**Father History**	
130<=x<140	-1 to 1	Yes	1
x>=140	1	No	-1
		Mother History	
		Yes	1
		No	-1
		Relative History	
		Yes	1
		No	-1

3.2.3 The proposed method for improving risk factors.

From experiments in section 4.2.2 (Result of experiment 2), it is shown that two new factors (alcohol and smoking) are useful. It also shows new appropriate ranges for traditional factors. Therefore, the proposed methodologies for improving risk factors are listed in the Table 5.

Table 5 The adjusted factors of diabetes risk that is improved from section 4.B.2

Traditional Factor	Range	New and Changed Factor			Range
Age (Years)		**WC**			
x<=44	-1	Male x<87 and Female x<80			-1
44<x<50	-1 to 1	Male x>=87 and Female x>=80			1
x>=50	1				
BMI		**FMH (from set E)**			
x<23	-1	**Father History**	**Mother History**	**Relative History**	
23<=x<=27	-1 to 1	Yes	Yes	Yes	1
x>27	1	No	No	No	-1
Gender		**Smoking**			
Female	-1	Yes			-1
Male	1	No			1
BP		**Alcohol/Week**			
x<130	-1	0<x<2			-1
130<=x<140	-1 to 1	x=0 or 2<=x<=3			0
x>=140	1	x>=4			1

3.3 Tuning BNN parameters

The objective of this section is to find appropriate values of the learning rate. From the experimental result in section 4.2.3 (Result of experiments 3), it was shown that the 0.004 learning rate is the best. More details can be found in section 4.2.3.

3.4 Diabetes prediction tools development

The development of a diabetes prediction tool is based on parameters modeled in Experiment 4, which is the web application using PHP programing. The web server is Apache Appserv and add-on software for the prediction process using Matlab (version 2011a).

In Figure 3, the workflow process will be started when a client enters the URL on a web browser. The input form will be presented as shown in Figure 1 for requesting personal data. When the client fills in the input boxes, it will be converted into the proposed ranges that are proven from the Experiments 1-3. After the diagnosis is started, the range values will be passed to the web server for generating a Matlab command. Then, the command will be processed by Matlab for executing the calculation parameter file (.mat), which is the trained parameter model in the Experiment 4. The process will send the result to the web server to make a decision by initialized threshold (0.22 is the best performing base on Set Proposed in the Table 8). Finally, the web server will respond and trigger the prediction result on presentation tier.

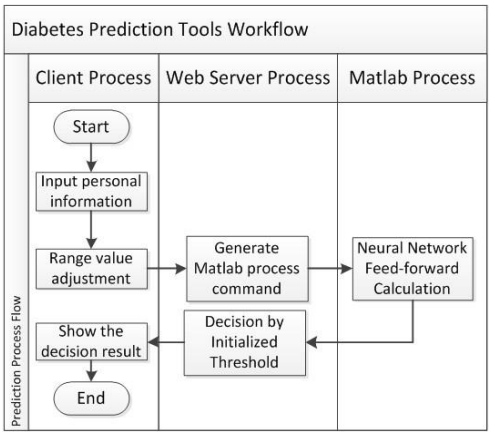

Figure 3 A workflow of the diabetes prediction tool

4. The experimental results

This section was separated into 2 major parts: experimental configuration and experimental results. The experimental configuration and results can be divided into 4 subsections.

For initializations of the weights and bias of BNN, 5 sets were created by random values between -0.5 to 0.5 before training started and were used for all experiments. This means 5 rounds of training for all sets by changing the initialization of weights and bias.

More learning rate values and initializations of the weights and bias increase the performance and avoid the error surface problem. The other is 60,000 epochs of learning that were used for all training in this paper.

4.1 Experiment

This research was divided into 2 groups with 80 percent training and 20 percent testing. Additionally, we used K-fold cross validation for training and testing in section 4.2.2 (Result of experiment 2) and 4.2.3 (Result of experiment 3). This technique was useful for assessing how the results of a statistical analysis will generalize to an independent data set. It was divided randomly into K groups. We defined K as 5 which randomly divided the data into five equal groups, meaning it was divided into 400 cases per group. The first fold used the first group for testing and other are training groups. Next, testing groups were changed to be other groups until completed. After that, the results were averaged for comparison of each training set.

The sample data is from the Bangkok Hospital during 2010 to 2012. In total, there are 2,000 Thai patient cases, containing 1,140 patients for Type 2 diabetes and 860 healthy persons.

4.1.1 Experiment 1

The targets of this experiment are to evaluate an appropriate number of hidden nodes and sequence of proving process for the proposed risk factors on diabetes diagnosis. Set A is a baseline (Kittiphol Wisaeng et al., 2009). Sets B to J are adjusted one by one of each proposed risk factors from Tables 3 and 4.

4.1.2 Experiment 2

For this experiment, the goal is to integrate the proposed factors by BNN and observe values of RMSE. To simplify, each proposed factors is integrated by sorting values RMSE (at number of hidden nodes equal to 50 as displayed in Figure 1) from the lowest to highest. This means that the RMSE of FMH factor is equal to 0.495 (the lowest) as shown in Figure 3, is firstly brought to BNN. The last one is RMSE of WC factor is equal to 0.5721 (the highest).

4.1.3 Experiment 3

The goal of this experiment is to tune BNN parameter by using the learning rate. To simplify, values of the learning rate are decreased by dividing by three as listed in Table 7.

4.1.4 Experiment 4

The development of prediction tools must use only one neural network structure. This experiment was the summarization of three experiments above (Experiments 1-3) for preparing the neural network structure that introduced the best performance for the prediction application. All sample data are used for the neural network training. The training used all of sample cases (2000 cases) without using K-Fold cross validation.

4.2 Results

For the performance evaluation criteria, this paper used RMSE for comparing the BNN structure, which indicated the ability of neural networks to predict and evaluate risk factors. The RMSE gave a relatively high weight to the large errors. RMSE is most useful when large errors are particularly undesirable.

4.2.1 Result of experiment 1

From Figure 4, it was found that an appropriate number of the hidden nodes are 50, taking the lowest mean RMSE with 0.54038. It was also shown that the proposed risk factors can be included in the diabetes diagnosis. Sets A to J at the appropriate number of hidden nodes (50) gave values of RMSE equal to 0.5617, 0.6046, 0.544, 0.5135, 0.495, 0.5721, 0.5683, 0.5401, 0.5041, and 0.5004, respectively.

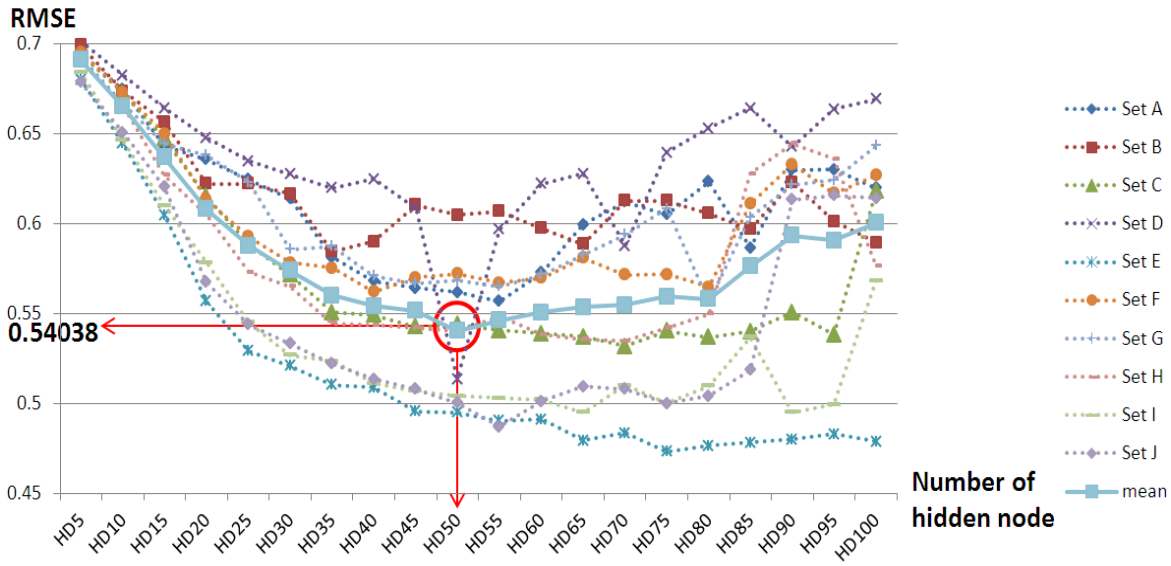

Figure 4 The number of hidden nodes at 50 introduces the lowest mean of RMSE as a marked point

4.2.2 Result of experiment 2

Table 6 lists values of RMSE (also shown in Figure 4) at number of hidden node equal to 50 in sequence from the lowest to highest, which proved the proposed factors (in 3rd column). It was found that sets E, I and J gave values of RMSE as 0.71782, 0.71626, and 0.71576, which are useful. However, the RMSE of set D was equal to 0.72108 taking a higher value than sets I and J. This degraded the results and was therefore removed from future experiments. Set H gave a lower RMSE than the sets I and J, and was acceptable. Sets B and C gave a higher RMSE and were therefore removed. Finally, sets F and G introduced a lower RMSE and were subsequently included.

Table 6 The integration sequence for proving each of the proposed risk factors

Process Sequence	Set Adjusting	Adjusting Factor	RMSE	Accuracy
1	A	-	0.72408	83.00%
2	E	FMH	0.71782	83.35%
3	I and J	alcohol	0.71626 and .71576	83.35% and 83.50%
4	D	BP	0.72108	83.55%
5	H	Smoking	0.7138	83.60%
6	B and C	BMI	0.71558 and 0.71722	83.65% and 83.00%
7	F and G	WC	0.71424 and 0.71378	83.20% and 83.65%

4.2.3 Result of experiment 3

In Table 7, set number five gave RMSE that are 0.7551, 0.6818, 0.7087, 0.6717, and 0.7333. The result of tuning reduced the value of RMSE as listed in Table 6 rows 7th from 0.71378 to 0.71012 (an average of five RMSE value from cross validation as listed in the bottom row). It increased the value of accuracy from 83.65% to 84.00%. Additionally, it was shown that values of lower learning rate introduced a lower RMSE. After the epochs reached a minimal point, additional epochs cannot give a higher performance in all cases.

Table 7 The results of the BNN tuning that is obtained from adjusting input set in Table 6

Set No.	Learning Rate	Epochs	Minimum of RMSE	Maximum of Accuracy
1	0.3	300	0.72716	83.00%
		600	0.72716	83.00%
		60,000	0.72716	83.00%
2	0.3, 0.1	300	0.71384	83.00%
		900	0.71378	83.13%
		60,000	0.71378	83.13%
3	0.3, 0.1, 0.033	1,200	0.71132	83.88%
		1,500	0.7113	84.00%
		60,000	0.7113	84.00%
4	0.3, 0.1, 0.033, 0.011	3,000	0.71086	83.88%
		4,500	0.71058	84.00%
		60,000	0.71058	84.00%
5	0.3, 0.1, 0.033, 0.011,0.004	3,000	0.7112	83.75%
		10,500	0.71012	84.00%
		60,000	0.71012	84.00%

4.2.4 Result of experiment 4

The preparation of prediction tools is a summarization of factors and neural network parameter model, which gave the highest performance. In Table 8, the comparison of the factor models shows that the new factor model (Set Tuning) enhances the performance better than baseline model result (Set A). The weight and bias parameters of neural network that gave the lowest value of RMSE are the foundation of the prediction tools.

Table 8 The result comparison of baseline and new model

Set Name	Learning Rate	Epochs	Minimum of RMSE
Set A (baseline)	0.3, 0.1, 0.033, 0.011,0.004	57,000	0.5639
		58,500	0.5628
		60,000	0.5617
Set Proposed	0.3, 0.1, 0.033, 0.011,0.004	57,000	0.4189
		58,500	0.418
		60,000	0.4171

5. Conclusion

The paper presents the new Type 2 diabetes risk factors of alcohol consumption in U-shape and smoking. Additionally, the new range of the traditional factor WC is proposed (cut point: 87 cm for males and 80 cm for females). The sample data was collected from the Bangkok Hospital during 2010 to 2012 consisting of 2,000 Thai people cases, of which 1,140 were patients for Type 2 diabetes and 860 were healthy volunteers. A predicable accuracy of the previous proposed risk factors model was 84.00 %, which it is better than the baseline up to 1.20 %. In this paper, the new proposed factors model introduces a 25.75 % better performance in RMSE than the baseline factors model compared with the baseline. Finally, the appropriate model is implemented to be the diabetes prediction tool. The application program is based on PHP web application and works in conjunction with MATLAB (2011a) for predicting calculations. In future work, the diabetes prediction application will be able to real-time self-learn from actual diabetes diagnoses for continually improving performance. However, there may be still other diabetes risk factors that have not been mentioned in this paper.

6. References

Aekphakorn W. (2005). *Diabetes risk score*. Bangkok, Thailand: Office of Health Information System. -In Thai-[วิชัย เอกพลากร (Wichai Aekphakorn). "การศึกษาพัฒนาดัชนีความเสี่ยงตอเบาหวาน" กรุงเทพฯ : สำนักงานพัฒนาระบบข้อมูลข่าวสารสุขภาพ, 2548.]

Baan, C. A., Ruige, J. B., Stolk, R. P., Witteman, J. C., Dekker, J. M., Heine, R. J., & Feskens, E.J. (1999). Performance of a

predictive model to identify undiagnosed diabetes in a health care setting. *Diabetes Care*, *22*, 213-219.

Baliunas, D. O., Taylor, B. J., Irving, H., Roerecke, M., Patra, J., Mohapatra, S., & Rehm, J. (2009). Alcohol as a risk factor for Type 2 diabetes. *Diabetes Care*, *32*, 2123-2132.

Barlow-Stewart, K. (2007). *The Australasian Genetics Resource Book*. Sydney, Australia: The Centre for Genetics Education.

Centers for Disease Control and Prevention. (2007). Types of diabetes. *National Diabetes Fact Sheet*, 2007.

Clair, C., Bitton, A., Meigs, James B., & Rigotti, N.A. (2011). Relationships of cotinine and self-reported cigarette smoking with hemoglobin A1c in the U.S. *Diabetes Care*, *34*, 2250-2255.

Glümer, C., Carstensen, B., Sandbaek, A., Lauritzen, T., Jørgensen, T., & Borch-Johnsen, K. (2004). A Danish Diabetes Risk Score for targeted screening. *Diabetes Care*, *27*, 727-733.

Golden, S. H., Wang, N. Y., Klag, M. J., Meoni, L. A., & Brancati, F. L. (2003). Blood pressure in young adulthood and the risk of type 2 diabetes in middle age, *Diabetes Care*, *26*, 1110-1115.

Joslin Diabetes Center, BMI Calculator, (online). http://aadi.joslin.org/content/bmi-calculator, February 4, 2012.

Koppes, L. L., Dekker, J. M., Hendriks, H. F., Bouter, L M., & Heine, R. J. (2005). Moderate alcohol consumption lowers the risk of Type 2 diabetes. *Diabetes Care*, *28*, 719-725.

Luangruangrong W., Rodtook, A., & Chimmanee S. (2012). Study of Type 2 diabetes risk factors using neural network for Thai people and tuning neural network parameters, IEEE SMC 2012, Seoul, Korea.

Matoba, Y., Inoguchi, T., Nasu, S., Suzuki, S., Yanase, T., Nawata, H., & Takayanagi, R. (2007). Optimal cut points of waist circumference for the clinical diagnosis of metabolic syndrome in the Japanese population. *Diabetes Care*, *31*, 590-592.

Matsuzawa Y. (2005). Metabolic syndrome-definition and diagnostic criteria in Japan. *Atherosclerosis and Thrombosis*, *12*, 301.

Mohan, V., Deepa, R., Deepa, M., Somannavar, S., & Datta M. (2005). A simplified Indian Diabetes Risk Score for screening for undiagnosed diabetic subjects. *Association of Physicians of India*, *53*, 759-763.

Rimm, E. B., Manson, J. E., Stampfer, M. J., Colditz, G. A., Willett, W. C., Rosner, B., Hennekens, C. H., & Speizer, F. E. (1993). Cigarette smoking and the risk of diabetes in women. *American Journal of Public Health*, *83*, 211-214.

Rimm, E. B., Chan, J., Stampfer, M. J., Colditz, G. A., & Willett, W. C. (1995). Prospective study of cigarette smoking, alcohol use and the risk of diabetes in men. *British Medical Journal*. *310*, 555.

Schmidt, M. I., Duncan, B. B., Bang, H., Pankow, J. S., Ballantyne, C. M., Golden, S. H., Folsom, A. R. & Chambless, L. E. (2005). Identifying individuals at high risk for diabetes. *Diabetes Care*, *28*, 2013-2018.

Stern, M. P., Williams, K., González-Villalpando, C., Hunt, K. J., & Haffner, S. M. (2004). Does the metabolic syndrome improve identification of individuals at risk of Type 2 diabetes and/or cardiovascular disease?. *Diabetes Care*, *27*, 2676-2681.

Wilson, P. W., Meigs, J. B., Sullivan, L., Fox, C. S., Nathan, D. M., & D'Agostino, R. B., Sr. (2007). Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. *Annals of Internal Medicine*, *167*, 1068-1074.

Wisaeng K., Chiewchanwattana S., & Khamron Sunat. (2009). Risk factor analysis of diabetes mellitus diagnosis, A thesis for the degree of Master of Computer Science, Khonkaen University, 798-805.

World Health Organization, BMI classification, (online). http://apps.who.int/bmi/index.jsp?introPage=intro_3.html, February 4, 2012.