Original Article

Healthc Inform Res. 2017 July;23(3):169-175.
https://doi.org/10.4258/hir.2017.23.3.169
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Statistics and Deep Belief Network–Based Cardiovascular Risk Prediction

Jaekwon Kim, MS[1], **Ungu Kang, PhD**[2], **Youngho Lee, PhD**[2]

[1]Department of Computer and Information Engineering, Inha University, Incheon, Korea; [2]IT Department, Gachon University, Seongnam, Korea

Objectives: Cardiovascular predictions are related to patients' quality of life and health. Therefore, a risk prediction model for cardiovascular conditions is needed. **Methods:** In this paper, we propose a cardiovascular disease prediction model using the sixth Korea National Health and Nutrition Examination Survey (KNHANES-VI) 2013 dataset to analyze cardiovascular-related health data. First, statistical analysis was performed to find variables related to cardiovascular disease using health data related to cardiovascular disease. Second, a model of cardiovascular risk prediction by learning based on the deep belief network (DBN) was developed. **Results:** The proposed statistical DBN-based prediction model showed accuracy and an ROC curve of 83.9% and 0.790, respectively. Thus, the proposed statistical DBN performed better than other prediction algorithms. **Conclusions:** The DBN proposed in this study appears to be effective in predicting cardiovascular risk and, in particular, is expected to be applicable to the prediction of cardiovascular disease in Koreans.

Keywords: Cardiovascular Diseases, Deep Belief Network, Machine Learning, Cardiovascular Risk Prediction, KNHANES

I. Introduction

Cardiovascular diseases include hyperlipidemia, myocardial infarction, and angina pectoris. Cardiovascular disease is diagnosed by electrocardiography, ultrasound, blood tests, angiography, and so on. These methods are time-consuming and costly because they require many different tests. Recently, a cardiovascular disease prediction technique using machine learning has been developed to replace these diagnostic methods [1].

Medical IT combined with machine learning technology has increased the accuracy of disease prediction using predictive models generated from disease-related learning data [2]. However, since complex data is analyzed, a deep learning technique is required [3,4].

Many studies have been conducted on cardiovascular disease using machine learning. Khatib and Montazer [5] developed a heart disease risk prediction model based on the Dempster-Shafer evidence theory by designing a fuzzy-evidential hybrid inference engine. Krishnaiah et al. [6] developed a cardiovascular risk prediction system using fuzzy K-nearest neighbor (K-NN) classifiers for measured values to remove uncertainty. However, research on a prediction model for domestic cardiovascular disease is lacking [7,8].

In recent years, attention has focused on how to construct a prediction model based on big data and the development of deep learning technology.

Prediction models are based on artificial intelligence (AI), and many methods using machine learning, data mining,

Submitted: March 24, 2017
Revised: May 8, 2017
Accepted: May 18, 2017

Corresponding Author
Youngho Lee, PhD
IT Department, Gachon University, 1342 Seongnam-daero, Sujeong-gu, Seongnam13120, Korea. Tel: +82-31-750-4759, E-mail: lyh@gachon.ac.kr

databases, and statistics have been proposed [9]. Prediction models using these cutting-edge techniques have been used in many fields, and their value in the medical industry is gradually increasing.

A deep belief network (DBN) is an advanced learning method using artificial neural networks which involves a high level of technology and performs well [10]. A DBN consists of several layers of controlling restricted Boltzmann machine (RBM). It then performs supervised learning using backpropagation after unsupervised learning [11]. DBN has broad applications in various medical fields and is widely used for medical research because it performs well [12-14].

In this paper, we propose a cardiovascular disease prediction model. The sixth Korea National Health and Nutrition Examination Survey (KNHANES-VI) 2013 data set [15] was used to find cardiovascular-related health data. First, statistical analysis was performed to find variables related to cardiovascular disease using health data related to cardiovascular disease. Second, a model of cardiovascular risk prediction by learning based on the DBN was developed. Thus, variables were selected using statistical techniques, and learning with DBN was conducted using the selected variables.

The remainder of this article is structured as follows. Section II describes the dataset and proposes the method. Section III outlines the system implementation and compares its ability to discriminate cardiovascular risk and probability tables. Finally, Section IV presents conclusions and specifies further directions for future research.

II. Methods

The research structure of this study is presented in Figure 1. First, the dataset was defined and data was preprocessed. Second, the dataset was statistically analyzed. We uses statistical techniques to select variables to be used for learning. The analyzed dataset was divided into a training set (70%) and a testing set (30%) (Table 1). Third, learning was done based on the DBN using the training set. Finally, the performance of the cardiovascular prediction model generated through learning was measured.

1. Dataset

The KNHANES-VI contains data from the Korea Centers for Disease Control and Prevention. KNHANES identifies the health and nutritional status of the population and selects the vulnerable groups that must be prioritized to calculate the statistics necessary to assess whether health policies and projects are being effectively delivered. It also provides statistical data on smoking, drinking, physical activity, and obesity requested by the World Health Organization (WHO) and the Organization for Economic Cooperation and Development (OECD) [15].

The Framingham risk score (FRS) has been used as a standard guideline for predicting cardiovascular risk for 10 years. Therefore, the attributes in these guidelines were used as a reference for data extraction [16].

Input variables for learning included age, gender, total cholesterol, high-density lipoprotein (HDL), systolic blood pressure (SBP), diastolic blood pressure (DBP), smoking, and

Table 1. Training and testing dataset

	Low risk	High risk	Total
Training (70%)	2,135	837	2,972
Testing (30%)	912	360	1,272
Total	1,223	1,197	4,244

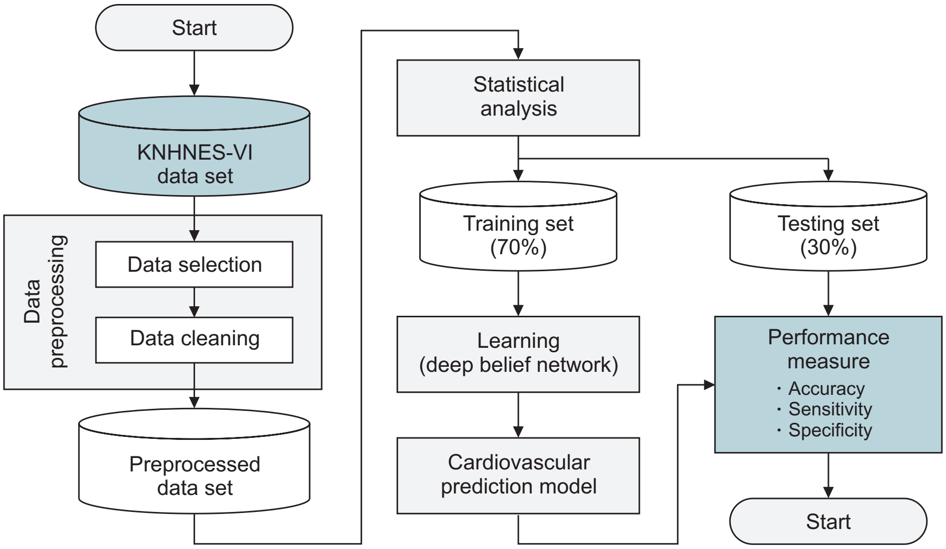


Figure 1. Study design.

https://doi.org/10.4258/hir.2017.23.3.169

diabetes. Output variables included cardiovascular diseases: hypertension, hyperlipidemia, myocardial infarction, and angina pectoris.

There are 8,108 experimental records in KNHANES-VI. Of these, 2,474 were from uncertain (non-respondent, null value) respondents, while 1,390 were records of people less than 30 years old. The final dataset comprised 4,244 records. Figure 2 shows the data preprocessing procedure.

2. Statistical Analysis

The statistical techniques for feature selection used the nonparametric Mann-Whitney *U*-test and chi-square. The age, SBP, DBP, total cholesterol, and HDL cholesterol variables were analyzed using the *U*-test. Chi-square testing was used to analyze the gender, diabetes, and smoking variables. Here, any variable whose *p*-value was less than or equal to 0.05 was excluded.

IBM SPSS Statistics ver. 22.0 (IBM, Armonk, NY, USA) was used for statistical analysis. Several statistical analysis methods with several preoperative variables were compared to determine the most effective method to predict cardiovascular risk.

A confusion matrix and receiver operating characteristic (ROC) curve were used to compare predictive ability. A confusion matrix is a measure to evaluate the performance of the classifier. As shown in Figure 3, accuracy, sensitivity, and specificity were measured. The matrix was constructed for output variables (low risk, high risk) in the testing dataset for each analysis. The limit of significance for all tests was $p < 0.05$.

3. Deep Belief Network

A DBN is a deep learning technique that learns by composing multiple RBM layers. MATLAB R2016b was used for the DBN in this research. The DeepLearnToolbox by R. B. Palm was used for the DBN library [17].

The RBM, which is based on the Hopfield network, employs the energy function and obtains unit values probabilistically (using a Boltzmann distribution). The RBM is shown in Figure 4. The RBM consists of a visible unit layer and a hidden unit layer, and its internal connection intensity is 0 [18]. The DBN, in which structures of the RBM are connected sequentially, is shown in Figure 5 [10].

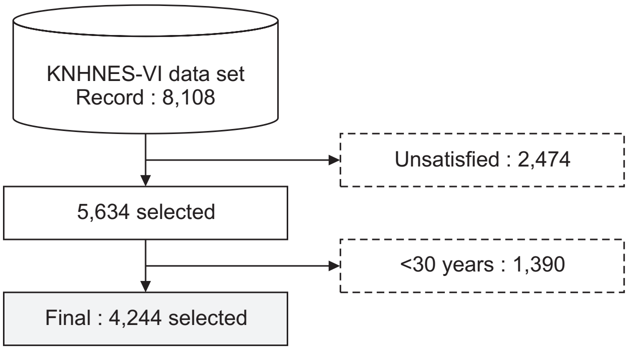
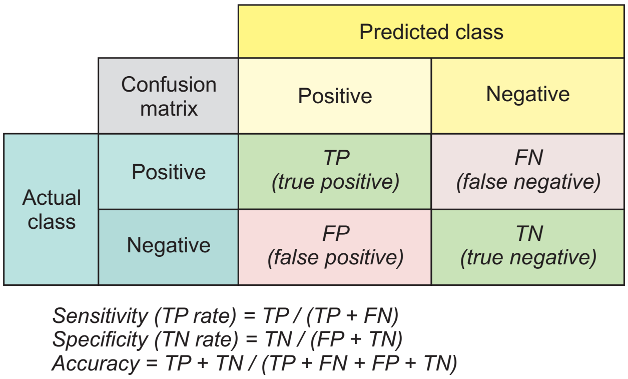


Figure 2. Data preprocessing.



	Confusion matrix	Predicted class	
		Positive	Negative
Actual class	Positive	*TP (true positive)*	*FN (false negative)*
	Negative	*FP (false positive)*	*TN (true negative)*

Sensitivity (TP rate) = TP / (TP + FN)
Specificity (TN rate) = TN / (FP + TN)
Accuracy = TP + TN / (TP + FN + FP + TN)

Figure 3. Confusion matrix.

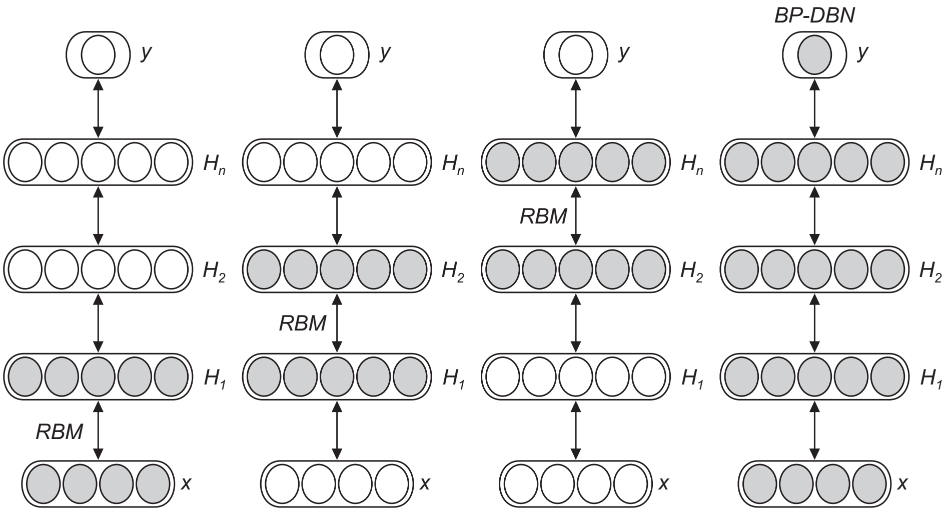


Figure 4. Restricted Boltzmann machine.

In the structure, the forefront hidden unit layer acts as the previous visible unit layer. DBN learning is done by configuring the visible layer and hidden layer 1 into a single RBM [19]. Once learning is complete, hidden layers 1 and 2 are trained via the RBM by giving a new input as a value of hidden layer 1. As such, learning is sequential up to the last layer.

A supervised learning-based classification technique using the DBN is the back propagation algorithm, which is configured in the uppermost layer in the DBN [20]. A classification prediction model using the backpropagation-DBN was created for this paper.

III. Results

1. Dataset Characteristics

The distribution of preoperative parameters among low- and high-risk records is shown in Table 2. The *p*-value >0.05 were gender and total cholesterol. In other words, there were six variables related to cardiovascular disease risk.

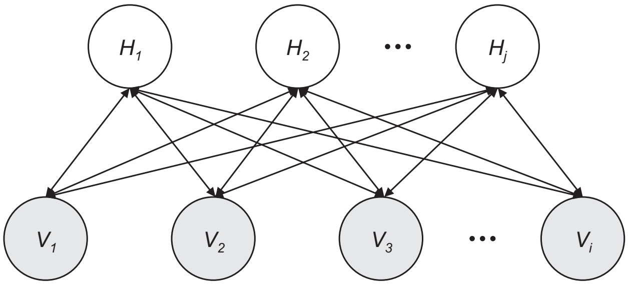


Figure 5. Deep belief network.

Table 2. Distribution of preoperative parameters among low- and high-risk records

	Low risk	High risk	*p*-value
Gender			0.594
Men	1,318	507	
Women	1,729	690	
Aver. age (yr)	48.57	62.78	0.000
Aver. SBP (mmHg)	115.54	127.61	0.000
Aver. DBP (mmHg)	75.49	76.28	0.030
Aver. total cholesterol (mg/dL)	191.28	189.89	0.260
Aver. HDL cholesterol (mg/dL)	52.64	49.70	0.000
Diabetes			0.000
No	2,936	918	
Yes	111	279	
Smoking			0.000
No	2,357	1,038	
Yes	690	159	

SBP: systolic blood pressure, DBP: diastolic blood pressure, HDL: high-density lipoprotein.

2. DBN Model

The DBN constructed a learning model using a training set. Six input variables (age, SBP, DBP, HDL, diabetes, smoking) and 1 output data were used. The DBN consisted of two steps. The first phase was the construction of the RBM network using unsupervised learning. The RBM settings were epoch, batch size, and momentum at 200, 12, and 0, respectively. In the second phase, the RBM network learned the backpropagation algorithm of supervised learning. The backpropagation options were epoch and batch size at 200 and 12, respectively.

The performance of the model differed depending on the number of nodes constituting the DBN. The error rate according to the number of nodes is shown in Table 3. Six nodes with one layer [4 8] showed the lowest error rate (0.2013). Therefore, it is best for construction of a DBN [4 8] (see Figure 6).

Table 3. Error rate according to number of nodes

	Error
1 Layer	
[1 0]	0.2830
[2 0]	0.2115
[3 0]	0.2830
[4 0]	0.2060
[5 0]	0.2075
[6 0]	0.2170
[7 0]	0.2138
[8 0]	0.2115
[9 0]	0.2083
[10 0]	0.2178
2 Layers	
[4 1]	0.2091
[4 2]	0.2830
[4 3]	0.2830
[4 4]	0.2028
[4 5]	0.2068
[4 6]	0.2288
[4 7]	0.2036
[4 8]	0.2013
[4 9]	0.2020
[4 10]	0.2028

https://doi.org/10.4258/hir.2017.23.3.169

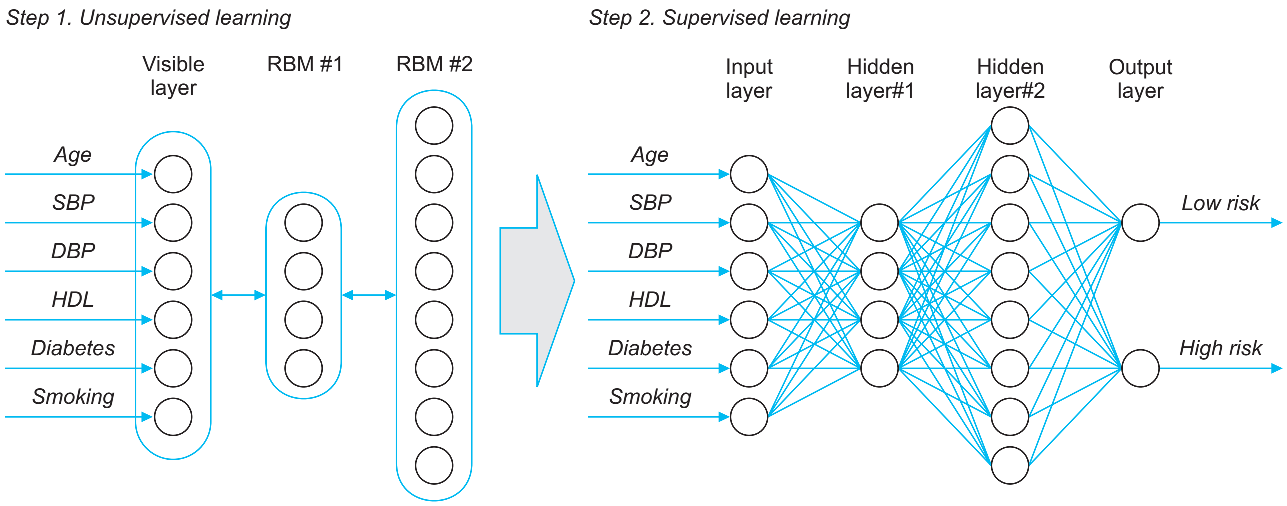


Figure 6. Deep belief network: (A) unsupervised learning, (B) supervised learning. SBP: systolic blood pressure, DBP: diastolic blood pressure, HDL: high-density lipoprotein.

Table 4. Confusion matrix results

Method	TP	FP	FN	TN
NB	788	124	141	219
LR	830	82	172	188
BPN	765	147	119	241
SVM	912	0	359	1
RF	794	118	172	188
DBN	900	12	305	55
Statistical DBN	823	89	116	244

TP: true positive, FP: false positive, FN: false negative, TN: true negative, NB: naïve Bayesian, LR: logistics regression, BPN: backpropagation network, SVM: support vector machine, RF: random forest, DBN: deep belief network.

Table 5. ROC curve results

	ROC curve	*p*-value	95% CI	
			Lower bound	Upper bound
NB	0.736 ± 0.017	0.000	0.703	0.769
LR	0.716 ± 0.018	0.000	0.682	0.751
BPN	0.701 ± 0.017	0.000	0.668	0.733
SVM	0.501 ± 0.018	0.938	0.466	0.537
RF	0.696 ± 0.018	0.000	0.662	0.731
DBN	0.570 ± 0.019	0.000	0.533	0.606
Statistical DBN	0.790 ± 0.016	0.000	0.759	0.821

ROC: receiver operating characteristic, CI: confidence interval, NB: naïve Bayesian, LR: logistics regression, BPN: backpropagation network, SVM: support vector machine, RF: random forest, DBN: deep belief network.

3. Experimental Results

We compared the performance of the proposed DBN with that of various machine learning techniques. The comparison models were naïve Bayesian (NB), logistics regression (LR), back propagation network (BPN), support vector machine (SVM), random forest (RF), DBN (using nine-variable input), and the proposed statistical DBN (using six-variable input) method. The confusion matrix results appear in Table 4. The ROC curve results are shown in Table 5.

Sensitivity, specificity, accuracy, and ROC curve results are shown in Figures 7–10.

Experimental results show that the proposed statistical DBN achieved the highest sensitivity, accuracy, and ROC curve performance. Specificity was 100% for SVM, and that for all others was low. In other words, SVM was effective in measuring low risk, but it could not predict important high

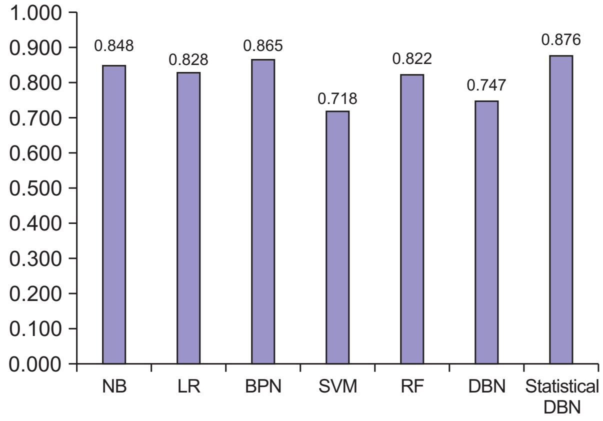


Figure 7. Sensitivity results.

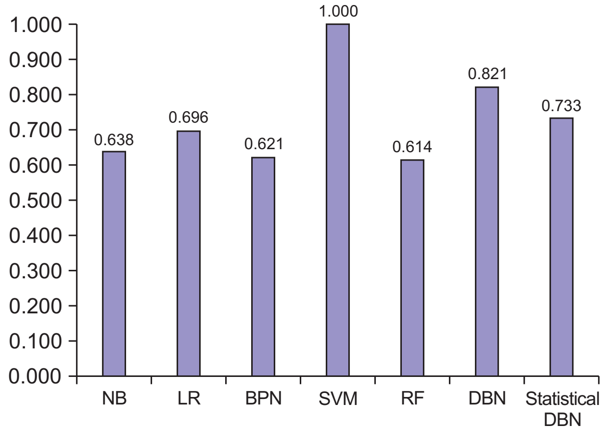


Figure 8. Specificity results.

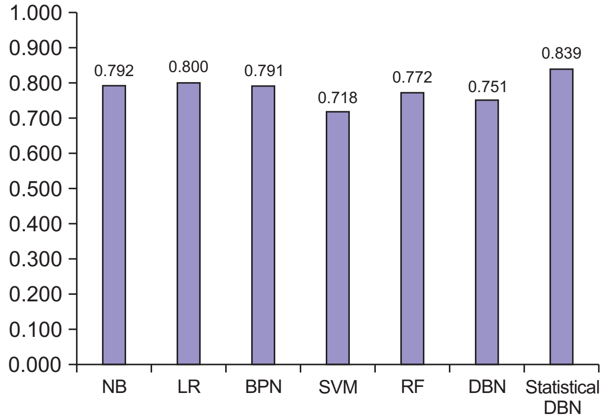


Figure 9. Accuracy results. NB: naïve Bayesian, LR: logistics regression, BPN: backpropagation network, SVM: support vector machine, RF: random forest, DBN: deep belief network.

risk. Sensitivity to measure high-risk prediction was highest at 87.6% for the proposed statistical DBN. Also, the existing DBN showed low performance because it does not consider unnecessary variables. It can be seen that unnecessary variables have a great influence on the measurement of cardiovascular disease. Therefore, the proposed model is able to achieve higher performance because it considers important variables.

IV. Discussion

This paper investigated methods that can be applied to predict the risk of cardiovascular disease. The existing methods for diagnosing cardiovascular disease are time-consuming and costly. However, cardiovascular disease risk can be predicted using various types of measured data when machine learning is applied.

In this paper, we implemented a risk prediction model using KNHANES-VI data. A DBN was used to implement the cardiovascular disease risk prediction model. Data analysis using statistical techniques showed that age, SBP, DBP, HDL, smoking, and diabetes were associated with cardiovascular risk. In other words, the prediction system can predict risk using six variables. The prediction model utilizes a DBN. The DBN consists of four input variables and one output variable. The experimental results show that it performed better than other methods. The method proposed in this paper appears to be effective for the risk prediction of cardiovascular disease and is expected to be particularly applicable to cardiovascular disease prediction in Koreans.

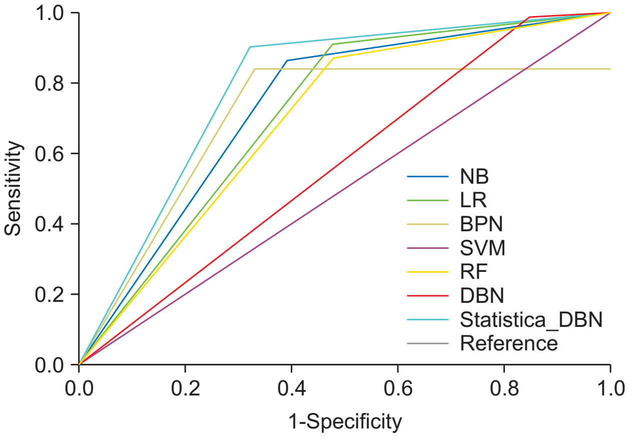


Figure 10. ROC curve result. NB: naïve Bayesian, LR: logistics regression, BPN: backpropagation network, SVM: support vector machine, RF: random forest, DBN: deep belief network.

Future research will focus on deep learning research to improve the performance of DBN node optimization and prediction.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

1. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J Clin Epidemiol 2013;66(4):398-407.
2. Song MH, Kim SH, Park DK, Lee YH. A multi-classifier based guideline sentence classification system. Healthc Inform Res 2011;17(4):224-31.
3. Tomar D, Agarwal S. Feature selection based least

square twin support vector machine for diagnosis of heart disease. Int J Biosci Biotechnol 2014;6(2):69-82.

4. Kim JK, Rho MJ, Lee JS, Park YH, Lee JY, Choi IY. (2016). Improved prediction of the pathologic stage of patient with prostate cancer using the CART-PSO optimization analysis in the Korean population. Technol Cancer Res Treat 2016 Dec 16 [Epub]. http://journals.sagepub.com/doi/abs/10.1177/1533034616681396.
5. Khatibi V, Montazer GA. A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment. Expert Syst Appl 2010;37(12):8536-42.
6. Krishnaiah V, Narsimha G, Chandra NS. Heart disease prediction system using data mining technique by fuzzy K-NN approach. In: Satapathy S, Govardhan A, Raju K, Mandal J, editors. Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1. Cham: Springer International Publishing; 2015. pp. 371-84.
7. Lee DY, Rhee EJ, Choi ES, Kim JH, Won JC, Park CY, et al. Comparison of the predictability of cardiovascular disease risk according to different metabolic syndrome criteria of American Heart Association/National Heart, Lung, and Blood Institute and International Diabetes Federation in Korean men. Korean Diabetes J 2008; 32(4):317-27.
8. Kim JK, Lee JS, Park DK, Lim YS, Lee YH, Jung EY. Adaptive mining prediction model for content recommendation to coronary heart disease patients. Clust Comput 2014;17(3):881-91.
9. Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Ithaca (NY): arXiv.org; c2017 [cited at 2017 Jul 1]. Available: https://arxiv.org/abs/1702.05747.
10. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Comput 2006;18(7): 1527-54.
11. Dahl GE, Yu D, Deng L, Acero A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans Audio Speech Lang Process 2012;20(1):30-42.
12. Abdel-Zaher AM, Eldeib AM. Breast cancer classification using deep belief networks. Expert Syst Appl 2016; 46:139-44.
13. Tamilselvan P, Wang P. Failure diagnosis using deep belief learning based health state classification. Reliab Eng Syst Saf 2013;115:124-35.
14. Fakoor R, Ladhak F, Nazi A, Huber M. Using deep learning to enhance cancer diagnosis and classification. Proceedings of the International Conference on Machine Learning; 2013 Jun 17-19; Atlanta, GA.
15. Korea Center for Disease Control and Prevention. The sixth Korea National Health & Nutrition Examination Survey (KNHANES-VI) 2013 [Internet]. Cheongju: Korea Center for Disease Control and Prevention; c2017 [cited at 2017 Jul 1]. Available: http://knhanes.cdc.go.kr/.
16. Ankle Brachial Index Collaboration, Fowkes FG, Murray GD, Butcher I, Heald CL, Lee RJ, et al. Ankle brachial index combined with Framingham Risk Score to predict cardiovascular events and mortality: a meta-analysis. JAMA 2008;300(2):197-208.
17. Palm RB. DeepLearnToolbox: a MATLAB toolbox for deep learning [Internet]. San Francisco (CA): GitHub Inc.; c2017 [cited at 2017 Jul 1]. Available: https://github.com/rasmusbergpalm/DeepLearnToolbox.
18. Hinton GE. A practical guide to training restricted Boltzmann machines. Toronto: University of Toronto; 2010.
19. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006; 313(5786):504-7.
20. Salama MA, Hassanien AE, Fahmy AA. Deep belief network for clustering and classification of a continuous data. Proceedings of 2010 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT); 2010 Dec 15-18; Luxor, Egypt. p. 473-7.