Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases

Berina Alić
Genetics and Bioengineering
International Burch University
Sarajevo, Bosnia and Herzegovina
berina_alic@hotmail.com

Lejla Gurbeta[1,2], Almir Badnjević[1,2,3]
[1]Verlab Ltd. Sarajevo
[2]International Burch University, Sarajevo
[3]Technical faculty Bihac, University of Bihac
lejla@verlab.ba, almir@verlab.ba

Abstract— **This paper presents the overview of machine learning techniques in classification of diabetes and cardiovascular diseases (CVD) using Artificial Neural Networks (ANNs) and Bayesian Networks (BNs). The comparative analysis was performed on selected papers that are published in the period from 2008 to 2017. The most commonly used type of ANN in selected papers is multilayer feedforward neural network with Levenberg-Marquardt learning algorithm. On the other hand, the most commonly used type of BN is Naïve Bayesian network which shown the highest accuracy values for classification of diabetes and CVD, 99.51% and 97.92% retrospectively. Moreover, the calculation of mean accuracy of observed networks has shown better results using ANN, which indicates that higher possibility to obtain more accurate results in diabetes and/or CVD classification is when it is applied to ANN.**

Keywords-machine learning; diabetes; cardiovascular disease; Artificial Neural Network, Bayesian Network

I. INTRODUCTION

Machine learning (ML) is subfield of Artificial Intelligence that solves the real world problems by "providing learning ability to computer without additional programming" [1]. The machine learning has developed from the efforts of researching whether computers could gather knowledge to mimic the human brain. The first attempts of ML were in 1952 when Arthur Samuel developed the first game-playing program for checkers, to accomplish enough skills to win against a world checker champion. Later in 1957, Frank Rosenblatt created an electronic device which has the ability to learn how to solve complex problems by imitating the process in human brain [1]. Development of ML contributed to the greater use of computers in medicine [2].

According to artificial intelligence market research firm 'TechEmergence' [3] and the researcher from the paper [4], the major machine learning applications in medicine are: smart electronic health records, drug discovery, biomedical signal processing and disease identification and diagnosis. In most cases of disease identification and diagnosis, the development of ML systems is considered as an attempt to imitate the medical experts' knowledge in the identification of disease. Since ML allows computer programs to learn from data developing a model to recognize common patterns and being able to make decisions based on gathered knowledge, it does not have difficulties with the incompleteness of used medical database [4]. In medical application, the most famous machine learning technique is classification because it corresponds to problems appearing in everyday life, among which the most usually applied techniques are Artificial Neural Networks (ANNs) and Bayesian Network (BNs).

The usage of machine learning in disease classification is very frequent [5-13] and scientists are even more interested in the development of such systems for easier tracking and diagnosis of diabetes and cardiovascular diseases. According to World Health Organization (WHO), both diabetes and cardiovascular disease (CVD) are among top ten causes of death worldwide [14]. The research from the January 2017 showed that the number one cause of death worldwide are CVDs. The world's biggest killer is taking the leading position in the list of top ten causes of deaths in the last 15 years and in 2015 was counting for 15 million deaths [15]. On the other hand, the first WHO Global report on diabetes demonstrated that in the period from 1980 to 2014, the number of adults with diabetes has risen from 108 million to 422 million, and the number of victims of diabetes in period from 2000 to 2015 increases from less than 1 million to 1.6 million people [16]. The morbidity and mortality from diabetes and CVD indicate the need for early classification of patients which can be achieved developing machine learning models. These models enable analysis of bigger and more complex data in order to achieve more accurate results and guide better decisions in real time without human intervention.

This study was designed to perform a review of Artificial Neural Network and Bayesian Network and their application in classification of diabetes and CVD diseases. The purpose is to show the comparison of these machine learning techniques and to discover the best option for achieving the highest output accuracy of the classification.

II. METHODS

This paper represents the comparison of application of two machine learning techniques, Artificial Neural Network and Bayesian Network in classification of diabetes and

978-1-5090-6742-8/17/$31.00 ©2017 IEEE

cardiovascular diseases. Guided by experience of researchers from the papers [17,18] that also reviewed machine learning techniques but in different field of studies, the literature review was done using 20 published papers in order to obtain the relevant results about diabetes and CVD classification in the period from 2008 to 2017.

Criteria for the paper selection were:

- the paper must be in English,
- published in the period 2008-2017,
- full text available,
- include classification of diabetes or CVD,
- disease classification by Artificial Neural Network or Bayesian Network and
- the results must indicate the accuracy of the network.

As it is presented in paper [19] each category compares results of 5 different papers. Thus, among 20 selected papers, 5 papers include classification of diabetes by ANN, 5 papers classification of CVD by ANN, 5 selected papers represents classification of diabetes using Bayesian Network and 5 papers shown CVD classification with Bayesian Network.

A. *Artificial Neural Network*

Artificial neural network uses supervised learning to classify input data into desired output. It consists of artificial neurons with weighted interconnections that modulate the effect of the associated input signals. The way how ANN uses supervised learning to classify input parameters of diabetes or CVD is shown in the Fig. 1.

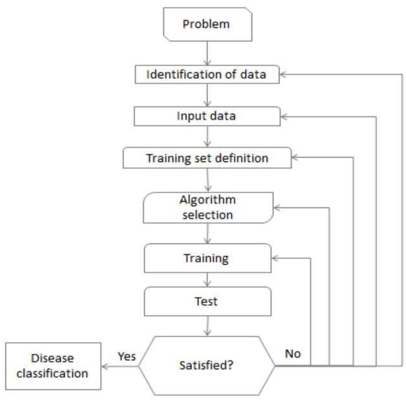

Figure 1. The process of classification input into desired output

The first step in classification of diabetes or CVD using ANN is to collect and identify data that will be used as an input to the network. The network is trained with defined training dataset and chosen training algorithm. After the training process, the ANN is additionally tested in order to obtain the feedback whether the network successfully classifies the disease.

Out of 20 selected papers, 10 papers show results of classification of diabetes and CVD using ANN. Table I represents the types of neural networks used for performing classification of mentioned diseases.

TABLE I. ANN TYPES FOR CLASSIFICATION OF DIABETES AND CVD

Paper	Type of ANN
DIABETES	
[20]	Multilayer feedforward neural network with sigmoid transfer function
[21]	Feedforward neural network using Levenberg-Marquardt method
[22]	Multilayer perceptron with backpropagation learning algorithm and genetic algorithm
[23]	Two-layer feedforward neural network with sigmoid function
[24]	Probabilistic neural network
CVD	
[25]	Multilayer neural network with statistical backpropagation of error
[26]	Backpropagation neural network with sigmoid transfer function
[27]	Feedforward neural networks with sigmoid transfer function using Levenberg -Marquardt learning algorithm and SCG
[28]	Feedforward multilayer perceptron with sigmoid activation function trained with backpropagation algorithm
[29]	MLP neural network with sigmoid transfer function

The overview of Artificial Neural Networks used for classification of diabetes and CVD (Table I) shows that the most commonly used type of network in both diseases is multilayer feedforward neural network. As training algorithm, most of authors of selected papers [17-26] have decided to use Levenberg-Marquardt learning algorithm. Each network uses error backpropagation algorithm to compare the system output to the desired output value, and uses the calculated error to direct the training. The difference in the architectures of these networks is in transfer function where sigmoid transfer function is the most commonly used one.

B. *Bayesian Network*

Bayesian networks (BNs) are probabilistic graphical models for reasoning under uncertainty. This model represents the set of random variables (discrete or continuous), where the arcs represent direct connections between them, and their conditional dependencies through directed acyclic graph [30]. The main reason why scientist have an interest in application of Bayesian Network in classification of diabetes or CVD [31-40] is that BN uses algorithms which are based on probability theory. This theorem is explicitly suitable for problems such as classification and regression [41].

Out of 20 selected papers, 10 papers show results of classification of diabetes and CVD using Bayesian Network. Table II represents the types of Bayesian networks used for performing classification of mentioned diseases.

TABLE II. BN TYPES FOR CLASSIFICATION OF DIABETES AND CVD

Paper	Type of BN
DIABETES	
[31]	Naïve Bayesian Network
[32]	Naïve Bayesian Network
[33]	Naïve Bayesian Network
[34]	MLP + Naïve Bayesian Network
[35]	Naïve Bayesian Network
CVD	
[36]	Markov blanket estimation
[37]	Dynamic Bayesian network
[38]	Naïve Bayesian network
[39]	Naïve Bayesian network
[40]	Naïve Bayesian network

The overview of Bayesian Networks used for classification of diabetes and CVD (Table II) shows that the most commonly used type of network in both diseases is Naïve Bayesian network. Naive Bayesian networks are very simple BNs which are composed of directed acyclic graphs with only one unobserved node and several observed nodes. This type of BNs applies Bayes' theorem with strong independence assumptions between features and does not need a long computational time for training which is its major advantage.

III. RESULTS

In the comparison of application of Artificial Neural Network and Bayesian Network for classification of diabetes and CVD, different values for the network accuracy have been achieved. Fig. 2 represents the results of trained ANN and BN for classification of diabetes from selected papers [20-24, 31-35]. On the left side, it can be seen that accuracy of diabetes classification using ANN varies between 72.2 and 99 %. On the right side, it is shown that accuracy of diabetes classification using BN varies between 71% and 99.51%.

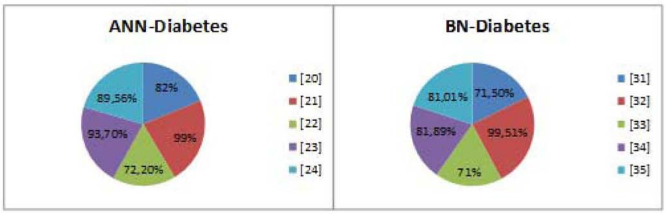

Figure 2. Accuracy for the classification of diabetes

According to compared results, the highest accuracy was achieved in Bayesian Network but also the smallest accuracy was shown in Bayesian Network. In order to obtain more information from this comparison, the Fig. 3 shows the accuracy values of both networks when they are lined from the lowest to the highest one (1-5).

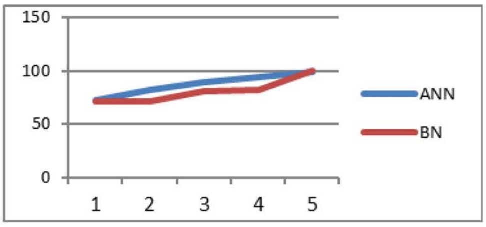

Figure 3. ANN and BN accuracy comparison

The Fig. 3 shows two curves where the blue stands for ANN and red stands for BN. Even though the highest accuracy belongs to BN, it is noticeable that the ANN curve in more cases shows higher accuracy than BN curve. Also, when we compare the mean accuracy of both, ANN and BN, we obtain ANN accuracy of 87.29% and BN accuracy of 80.98%, which indicates that there is higher possibility to achieve higher accuracy for diabetes classification when it is done by ANN.

Fig. 4 represents the results of trained ANN and BN for classification of CVD from selected papers [25-29, 36-40]. On the left side, it can be seen that accuracy of CVD classification using ANN varies between 80 and 95.91%. On the right side, it is shown that accuracy of CVD classification using BN varies between 78% and 97.92%.

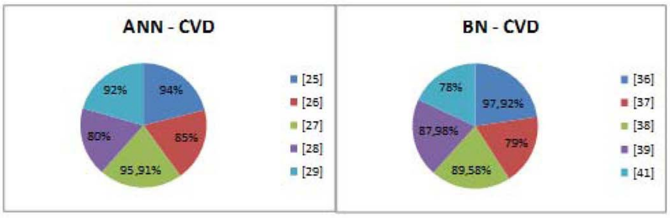

Figure 4. Accuracy for the classification of CVD

In accordance to Fig. 4, the highest accuracy was achieved in Bayesian Network as well as the smallest accuracy for the classification of CVD. For the better overview of this comparison, the accuracy values of both types of networks are lined from the lowest to the highest (1-5) and their values are represented on Fig. 5.

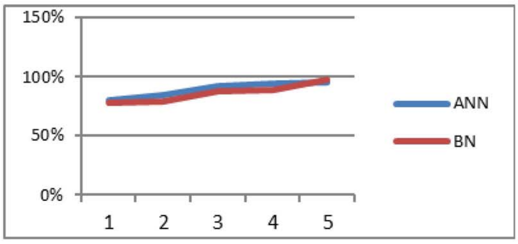

Figure 5. ANN and BN accuracy comparison

When these two curves are compared, it can be concluded that even though the highest accuracy is obtained in BN, in more cases (4 of 5) higher accuracy was achieved with ANN. Also, in the comparison of the mean accuracy values ANN accuracy of 89.38 % is higher than BN accuracy of 86.49%.

Moreover if we calculate population standard deviation (σ) using the equation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x-\bar{x})^2} \qquad (1)$$

the results shown in Table III will be obtained.

TABLE III. STANDARD DEVIATION FOR EACH CATEGORY

σ	ANN	BN
DIABETES	9,37	10,33
CVD	5,96	7,36

In this case, when we compare standard deviation of ANN and BN for both diabetes and CVD, in both cases the higher value is obtained when BN is used. This means that in BN we obtain higher deviation from mean value. Consequently, it can be concluded that it is higher possibility to achieve better accuracy and more reliable results for diabetes and CVD classification when it is performed by ANN.

IV. CONCLUSION

One of the biggest causes of death worldwide are diabetes and cardiovascular disease. The early classification of these diseases can be achieved developing machine learning models such as Artificial Neural Network and Bayesian Network. In comparison of mean accuracy of 10 scientific papers about diabetes classification and 10 papers about CVD classification it was concluded that the higher accuracy was achieved with ANN in both cases (87.29 for diabetes and 89.38 for CVD). The used Naïve Bayesian network, due to the assumption of independence among observed nodes, might be less accurate than ANN approach. So, in accordance to obtained result it can be concluded that the higher possibility to obtain better accuracy in classification diabetes and/or CVD is when it is applied to Artificial Neural Network.

References

[1] N. Sandhya, K.R. Charanjeet, A review on Machine Learning Techniques, International Journal on Recent and Innovation Trends in Computing and Communication, 2016, ISSN: 2321-8169, 395 – 399.

[2] A. Ghaheri, S. Shoar, M. Naderan and S.S. Hoseini, The applications of genetic algorithms in medicine. Oman medical journal, 2015, 30(6), 406.

[3] D. Fagella, 7 Applications of Machine Learning in Pharma and Medicine, 2017, Available at: https://goo.gl/1SlR5k.

[4] G.D. Magoulas and A. Prentza, Machine learning in medical applications. In Machine Learning and its applications, Springer Berlin Heidelberg, 2001, pp. 300-307).

[5] Badnjevic A, Cifrek M, Koruga D, Osmankovic D. „Neuro-fuzzy classification of asthma and chronic obstructive pulmonary disease“ BMC Medical Informatics and Decision Making Journal, 2015.

[6] Badnjevic A, Koruga D, Cifrek M, Smith HJ, Bego T. „Interpretation of pulmonary function test results in relation to asthma classification using integrated software suite“, IEEE MIPRO, pp: 140 -144, 2013. Croatia.

[7] Badnjevic A., Cifrek M., „Classification of asthma utilizing integrated software suite“, MBEC, 07-11. September 2014., Dubrovnik, Croatia

[8] Avdakovic S., Omerhodzic I., Badnjevic A., Boskovic D., „Diagnosis of Epilepsy from EEG Signals using Global Wavelet Power Spectrum“, MBEC, 07-11. September 2014., Dubrovnik, Croatia

[9] Badnjevic A, Gurbeta L, Cifrek M, Marjanovic D, „Classification of Asthma Using Artificial Neural Network“, IEEE MIPRO 2016.

[10] Aljovic A, Badnjevic A, Gurbeta L, „Artificial Neural Networks in the Discrimination of Alzheimer's disease Using Biomarkers Data“, IEEE MECO, 12 – 16 June 2016, Bar, Montenegro

[11] Alic B, Sejdinovic D, Gurbeta L, Badnjevic A, „Classification of Stress Recognition using Artificial Neural Network“, IEEE MECO, 2016.

[12] Fojnica A, Osmanovic A, Badnjevic A, „Dynamical Model of Tuberculosis-Multiple Strain Prediction based on Artificial Neural Network“, IEEE MECO, 2016, Bar, Montenegro

[13] Alic B, Gurbeta L, Badnjevic A, et.al., “Classification of Metabolic Syndrome patients using implemented Expert System”, CMBEBIH 2017. IFMBE Proceedings, vol 62. pp 601-607, Springer.

[14] World Health Organization, Top 10 Causes of Death, 2017, Available at: http://www.who.int/mediacentre/factsheets/fs310/en/.

[15] World Health Organization, Cardiovascular Disease, 2017, Available at: http://www.who.int/mediacentre/factsheets/fs317/en/

[16] World Health Organization, Diabetes, 2017, Available at: http://www.who.int/mediacentre/factsheets/fs312/en/

[17] Habibi, N., Hashim, S. Z. M., Norouzi, A., & Samian, M. R. (2014). A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli. BMC bioinformatics, 15(1), 134.

[18] Langarizadeh, M., & Moghbeli, F. (2016). Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review. Acta Informatica Medica, 24(5), 364.

[19] Fatima, M., & Pasha, M. (2017). Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications, 9(01), 1.

[20] Olaniyi, E. O., & Adnan, K. (2014). Onset diabetes diagnosis using Artificial neural network. International Journal of Scientific and Engineering Research, 5(10).

[21] Jayalakshmi, T., & Santhakumaran, A. (2010, February). A novel classification method for diagnosis of diabetes mellitus using artificial neural networks. DSDE, 159-163. (2010)

[22] Pradhan, M., & Sahu, R. K. (2011). Predict the onset of diabetes disease using Artificial Neural Network (ANN). International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004).

[23] Sejdinović, Dijana, et al. "Classification of Prediabetes and Type 2 Diabetes using Artificial Neural Network." Springer. CMBEBIH 2017.

[24] Soltani, Z., & Jafarian, A. (2016). A New Artificial Neural Networks Approach for Diagnosing Diabetes Disease Type II. International Journal of Advanced Computer Science & Applications, 1(7), 89-94.

[25] Atkov, O. Y., Gorokhova, S. G., Sboev, A. G., Generozov, E. V., Muraseyeva, E. V., Moroshkina, S. Y., & Cherniy, N. N. (2012). Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters. Journal of cardiology, 59(2), 190-194.

[26] Olaniyi, E. O., Oyedotun, O. K., & Adnan, K. (2015). Heart diseases diagnosis using neural networks arbitration. International Journal of Intelligent Systems and Applications, 7(12), 72.

[27] Colak, M. C. et al., Predicting coronary artery disease using different artificial neural network models/koroner arter hastaliginin degisik yapay sinir agi modelleri lie tahmini. The Anatolian Journal of Cardiology (Anadolu Kardiyoloji Dergisi), 8(4), 249-255, (2008).

[28] Can, M. (2013). Diagnosis of cardiovascular diseases by boosted neural networks.

[29] Sayad, A. T., & Halkarnikar, P. P. Diagnosis of heart disease using neural network approach. In Proceedings of IRF International Conference, 13th April-2014, Pune, India, ISBN (pp. 978-93).

[30] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

[31] Guo, Y., Bai, G., & Hu, Y. (2012, December). Using bayes network for prediction of type-2 diabetes. In Internet Technology And Secured Transactions, 2012 International Conference for (pp. 471-472). IEEE.

[32] Kumari, M., Vohra, R., & Arora, A. (2014). Prediction of Diabetes Using Bayesian Network.

[33] N. Sarma, S. Kumar, A.K. Saini, A Comparative Study on Decision Tree and Bayes Net Classifier for Predicting Diabetes Type 2, 2014, ISSN: 2278–0882, ICRTIET-2014.

[34] Dewangan L. A., & Agrawal, P. Classification of Diabetes Mellitus Using Machine Learning Techniques.

[35] Nai-arun, N., & Moungmai, R. (2015). Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science, 69, 132-142.

[36] Elsayad, A., & Fakr, M. (2015). Diagnosis of cardiovascular diseases with Bayesian classifiers. J. Comput. Sci., 11(2), 274-282.

[37] K. P. Exarchos, et al. Prediction of coronary atherosclerosis progression using dynamic Bayesian networks. IEEE EMBC, 2013.

[38] D.S. Medhekar, M.P. Bote & Deshmukh, S. D., Heart disease prediction system using naive bayes. Int. J. Enhanced Res. Sci. Technol. (2013).

[39] Patil, R. R., Heart disease prediction system using naïve bayes and jelinek-mercer smoothing. International Journal of Advanced Research in Computer Science and Communication Engineering, (2014).

[40] E. Miranda et al., Detection of CVD Risk's Level for Adults Using Naive Bayes Classifier. Healthcare Informatics Research, (2016).

[41] N. Sandhya and K.P. Charanjeet, A review on Machine Learning Techniques, 2016, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169, 395 – 399.