Performance comparison of Extreme Learning Machines and other machine learning methods on WBCD data set

Ömer Selim Keskin and Akif Durdu
Robotics Automation Control Laboratory (RAC-LAB), Electical-Electronics Engineering, Konya Technical University, Konya, Turkey
e208221001051@ktun.edu.tr, adurdu@ktun.edu.tr

Muhammet Fatih Aslan
Electrical-Electronics Engineering
Karamanoğlu Mehmetbey University
Karaman, Turkey
mfatihaslan@kmu.edu.tr

Abdullah Yusefi
Computer Engineering
Kabul University
Kabul, Afghanistan
a.yusefi@gmail.com

2021 29th Signal Processing and Communications Applications Conference (SIU) | 978-1-6654-3649-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/SIU53274.2021.9477984

***Abstract*— Breast cancer is one of the most common forms of cancer among women in our country and the world. Artificial intelligence studies are growing in order to reduce the mortality and early diagnosis needed for appropriate treatment. The Excessive Learning Machines (ELM) method, one of the machine learning approaches, is applied to the Wisconsin Breast Cancer Diagnostic (WBCD) dataset in this study, and the findings are compared to those of other machine learning methods. For this purpose, the same dataset is also classified using Multi-Layer Perceptron (MLP), Sequential Minimum Optimization (SMO), Decision Tree Learning (J48), Naive Bayes (NB), and K-Nearest Neighbor (KNN) methods. According to the results of the study, the ELM approach is more successful than other approaches on the WBCD dataset. It's also worth noting that as the number of neurons in the ELM grows, so does the learning ability of the network. However, after a certain number of neurons have passed, test performance begins to decline sharply. Finally, the ELM's performance is compared to the results of other studies in the literature.**

Keywords — breast cancer; machine learning; extreme learning machines; classification.

I. INTRODUCTION

Cancer is a disease that develops when cells multiply abnormally and uncontrollably. Breast cancer is caused by the uncontrolled proliferation of milk-producing cells in the breast [1]. Breast cancer is one of the most common cancers in women. "It is, for example, the most common cancer type in women in the United States and China. Since 2000, the incidence rate has been rising and every year, approximately 1.38 million new cases of breast cancer are diagnosed." Despite the development of numerous diagnostic and treatment techniques, breast cancer remains one of the top two cancers in terms of annual mortality rates in both countries [2]. Uncontrolled cell reproduction of 80% of women occurs in the mammary glands. Although breast cancer is thought to be a disease that only affects women, it can also affect men. In comparison to other cancer types, the incidence rate of breast cancer in men is about 1% [3]. It affects one out of every 100,000 men [4]. Breast cancer is diagnosed using a variety of tests, including mammography and clinical examinations. A breast cancer diagnosis is just as critical as other forms of cancer in terms of early detection. Although breast cancer is common, it is slow-growing cancer that can be detected early and treated successfully, lowering the mortality rate [5]. When looking at 5-year survival rates in our country, female cancer patients diagnosed in the early stages have a 90.0 percent survival rate [6].

Machine learning enables computers to quickly detect patterns in complex and large datasets using statistical, probabilistic, and optimization techniques. It is widely used in the detection and diagnosis of cancer due to these capabilities [7]. In the medical field, the accuracy (performance) percentage of machine learning methods is critical. This accuracy rate may reach the point of determining whether a person is sick or not, and if healthy people are incorrectly diagnosed, it can result in unnecessary harm to the person's body through drug injections or interventions. Therefore, a near-perfect success rate should be expected [1]. ELM also provides a high level of learning efficiency. Furthermore, with the non-repetitive training model, all parameters are set only once, allowing training to be completed quickly. Not only classification but also in other operations that many algorithms can do, ELM can perform better and faster [7]. Furthermore, unlike Artificial Neural Networks (ANN), they do not require classification parameter optimization. The general structure of ELM is depicted in Figure 1.

The GA-FL approach was used by Abonyi and Szeifert, who recorded a 95.57% success rate. According to Kim et al.'s study, this efficiency rate is 96.66% using their method. Moreover, Using the K-Nearest Neighbor (KNN) approach, which is commonly favored in the literature, Şahan et al. were able to achieve a 99.14% success rate. With the LS-SVM (Least Squares Support Vector Machines) method, Polat and

978-1-6654-3649-6/21/$31.00 ©2021 IEEE

Güneş were able to achieve a 98.53% success rate. In another work using the SVM model, Akay was able to achieve a 99.51% success rate. The ANN method was stated to have a success rate of 97.4% by Karabatak and Ince. Kahramanli and Allahverdi, on the other hand, found 99.31% progress in their analysis using the ANN method [8]. According to the literature review, each approach has advantages and disadvantages compared to each other. The ELM approach on the WBCD dataset was preferred in this study.

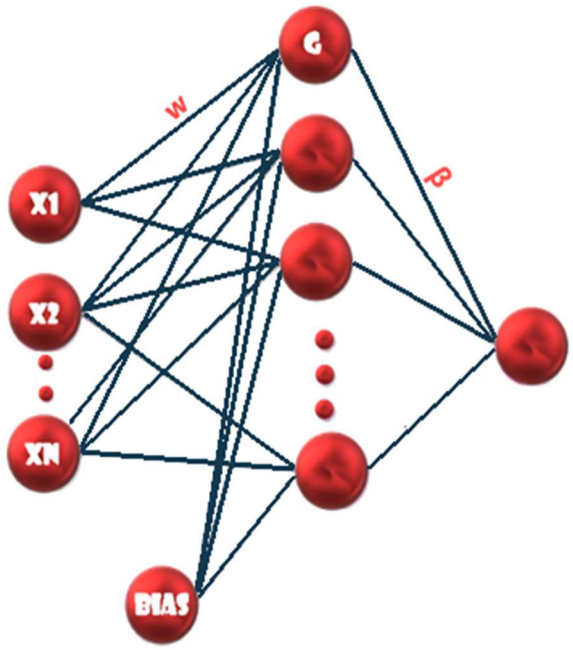

Fig. 1. ELM Structre

In this analysis, the methods of machine learning to be compared on the WBCD data set are described as ELM, Multi-Layer Perceptron (MLP), Sequential Minimal Optimization (SMO), Decision Tree Learning (J48), Naive Bayes (NB), and KNN. In comparison to the five methods, the ELM method generated a more reliable result in this analysis. The paper is structured as follows: In section II, it is explained how machine learning methods compare results, what the error matrix is, and what precision means. Additionally, the study's path has been shown. In section III, the results were presented, discussed, and graphically represented.

II. METHOD

In this study, 60% of the 569 total diagnoses in the Breast Cancer Wisconsin dataset were used for training and 40% for testing purposes, with the samples chosen at random. The classification output was evaluated over the test data after the training phase. It indicates the number of test data correctly predicted and the general classification accuracy is calculated by comparing the predicted classes with the ground truth classes of the test data.

A. *Performance Criteria*

The accuracy value is a ratio that indicates how many predictions machine learning models can correctly predict out of the total number of test data. It's calculated by dividing the number of correctly predicted values by the total number of predicted values. Equation 1 is used to measure the performance criterion, and these results are used to calculate accuracy values [9].

$$Accuracy = \frac{a+d}{a+b+c+d} \quad (1)$$

TABLE I. CONFUSION MATRİX

Accuracy Values		Prediction	
		Healthy	***Cancerous***
Correct	***Healthy***	62	1
	Cancerous	1	202

B. *Wisconsin* Breast Cancer Database *(WBCD) Dataset*

The WBCD dataset was created with the help of the UCI machine learning database [10]. It includes a dataset of 699 samples collected by Dr. WH Wolberg. There are a few missing elements in some of the examples. As a result, 16 samples were discarded. The remaining 683 data sets in the PAC model were tested after the removal of these 16 samples. WBCD has ten distinct features, which are described below. Variable coefficients of 1 to 10 are used in these properties. The target trait is coded as benign or malignant. The classification system uses "1" values for benign and "0" values for malignant conditions. The dataset developed by Dr. WH. Wolberg contains 444 benign cases and 239 malignant cases [11]. The 10 features which are measured directly in each cell nucleus and are of different types are as follows:

- Radius
- Texture
- Perimeter
- Area
- Smoothness
- Compactness
- Concavity
- Concave Points
- Symmetry
- Fractal Dimension

C. *Method of Application*

ELM, MLP, SMO, J48, NB, and KNN methods are some of the machine learning methods that have been compared in the literature. The accuracy value, which is measured over the

WBCD dataset and whose results are obtained, is calculated using these methods via Equation 1. In this analysis, 61% of the data was used for training. The number of hidden layers has also been set to 50. The ELM algorithm can be outlined in three steps as follows:

- The input weights (Wi) and the hidden layer threshold value (bi) are randomly generated.
- Hidden layer (H) output is calculated.
- Output weights are calculated according to Equation (2).

$$\beta = H^{+}Y.Y \tag{2}$$

A library has been created that is licensed under Riccardo Taormina's "GNU General Public License" and accessible through GitHub [13]. The ELM method was used to apply the classification process to the WBCD dataset with the aid of this library. The test accuracy rates obtained as a result of the analysis conducted were used to compare the outputs of the NB, MLP, J48, SMO, and KNN methods. Section III shows the results.

III. RESULTS

As a result, it was discovered that this research performed better than other machine learning approaches on the WBCD dataset while using the ELM method.

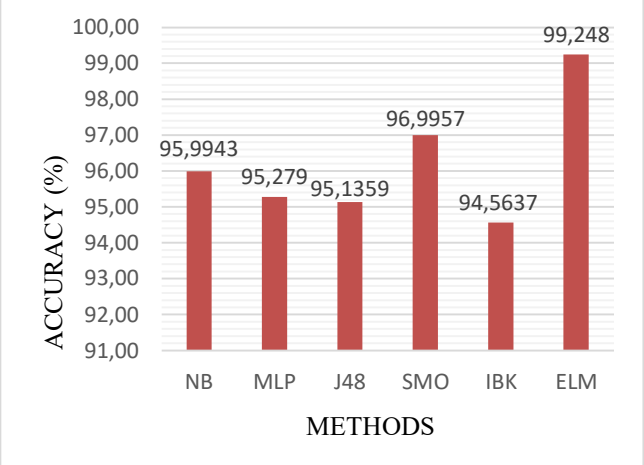

Fig. 2. Test accuracy result

The comparison of the accuracy value of the ELM method used in this study with other methods (NB, MLP, j48, SMO, IBK) is shown in figure 2. In cancer patients, 202 people's outcomes were correctly predicted, but one person's outcome was incorrectly predicted.

TABLE II. RESULT OF THE APPLICATION ERROR MATRIX

Accuracy Values		Prediction	
		Healthy	***Cancerous***
Correct	***Healthy***	62	1
	Cancerous	1	202

Figure 3 shows the sensitivity analysis on the number of hidden neurons based on the most recent experimental research. When looking at the graph in Figure 3, the training performance improves as the number of neurons increases, while the test performance drops sharply after a certain number of neurons.

In this study, the ELM method achieved a training accuracy of 96,403% and a test accuracy rate of 99,248% when the number of hidden layers is 50 and 60% of the data is used for training and 40% for testing.

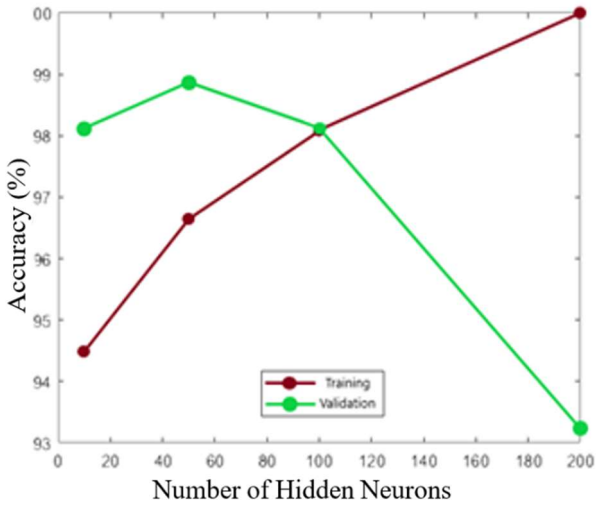

Fig. 3. Sensitivity analysis on the number of hidden neurons

In addition, when the number of hidden layers is 50 and 80% of the data is used for training and 20% for testing, an educational success of 97.258% and a test success of 99.265% are observed. If the rates are used for 90% education 10% test, 97.236% training and 98.529% test success has been observed.

ACKNOWLEDGMENT

Authors are grateful to the RAC-LAB, Turkey (www.rac-lab.com) for training and support. Furthermore; we are thankful for Mr. Süleyman Hanlı the R&D Unit Manager in Endüstriyel Elektrik Elektronik San. ve. Tic. Ltd. Şti., for his support. We wish him a good luck.

REFERENCES

[1] https://www.anadolusaglik.org/blog/meme-kanseri

[2] Chang-Min Kim, Roy C. Park, Ellen J. Hong, "Breast Mass Classification Using eLFA Algorithm Based on CRNN Deep Learning Model", Access 1IEEE, vol. 8, pp. 197312-197323, 2020.

[3] Gradishar WJ: Male breast cancer, in Harris JR, Lippman ME, Morrow M, Osborn CK (ed): Disease of the Breast. Philedelphia, Lippincott Williams and Wilkins,2000, pp 661-667

[4] Goss PE, Reid C, Pintilie M, et al. Male breast carsinoma: a review of 229 patients who presented to the Princess Margaret Hospital during 40 years: 1995-1996.Cancer. 1999 Feb 1;85(3):629-39.

[5] Çiğdem F., Ersin F., "The Effect of Women's Social Support and Self-Efficacy Perceptions on Early Diagnosis Behaviors of Breast Cancer" 16(3): 183-190, 2019.

[6] Türkiye Kanser İstatistikleri Ocak 2014. https://hsgm.saglik.gov.tr/depo/birimler/kanser-db/

[7] Cruz, J. A., Wishart, D. S., "Applications of machine learning in cancer prediction and prognosis" Cancer informatics, 2:59–77, 2006.

[8] Yılmaz KAYA, "A new intelligent classifier for breast cancer diagnosis based on a rough set and extreme learning machine: RS + ELM" Turk J Elec Eng & Comp Sci, 21: 2079 – 2091, 2013.

[9] Gouda I. Salama, M. B. Abdelhalim, and Magdy Abdelghany Zeid, "Breast Cancer Diagnosis on Three Different Datasets Using Multi Classifiers" International Journal of Computer and Information Technology (2277 – 0764) Vol.01– Issue 01, September 2012.

[10] UCI Repository of Machine Learning Databases. Available at https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/

[11] Sevli. O, "Göğüs Kanseri Teşhisinde Farklı Makine Öğrenmesi Tekniklerinin Performans Karşılaştırması" Avrupa Bilim ve Teknoloji Dergisi, (16), 176-185, 2019.

[12] https://github.com/rtaormina/ELM_MatlabClass