Comparison of Machine Learning Methods for Breast Cancer Diagnosis

Ebru Aydındag Bayrak, Pınar Kırcı
Department of Engineering Sciences
Istanbul University Cerrahpasa
stanbul, Turkey
ebruaydindag@gmail.com, pkirci@istanbul.edu.tr

Tolga Ensari
Department of Computer Engineering
Istanbul University Cerrahpasa
stanbul, Turkey
ensari@istanbul.edu.tr

***Abstract*— Cancer is the common problem for all people in the world with all types. Particularly, Breast Cancer is the most frequent disease as a cancer type for women. Therefore, any development for diagnosis and prediction of cancer disease is capital important for a healthy life. Machine learning techniques can make a huge contribute on the process of early diagnosis and prediction of cancer. In this paper, two of the most popular machine learning techniques have been used for classification of Wisconsin Breast Cancer (Original) dataset and the classification performance of these techniques have been compared with each other using the values of accuracy, precision, recall and ROC Area. The best performance has been obtained by Support Vector Machine technique with the highest accuracy.**

***Keywords*— machine learning; breast cancer; classification; early diagnosis.**

I. Introduction

Cancer is the second reason of human death all over the world and accounts for roughly 9.6 million deaths in 2018. Globally, for 1 human death in 6 can be said that is caused by cancer. Almost 70 percent of the deaths from cancer disease happen in countries that have low and middle income [1]. The most common cancer type among women are breast, lung and colorectal, which totally symbolize half of the all cancer cases. Also, breast cancer is responsible for the thirty percent of all new cancer diagnoses in women [2]. Machine learning (ML) methods ensure analyzing the data and extracting key characteristics of relationships and information from dataset. Also, it creates a computational model for best description of the data. Especially, according to in researches about cancer disease, it can be said that ML techniques can be handled on early detection and prognosis of cancer [3]. Asri et al. have compared some machine learning algorithms for the risk prediction and diagnosis of breast cancer. Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Naive Bayes (NB) and Decision Tree (C4.5) have been applied Wisconsin Breast Cancer (Original) dataset. SVM classification method has been given the highest accuracy value (97.13 %) with least error rate when the experimental results were compared [4].

Bazazeh and Shubair have investigated the comparative study of machine learning techniques as Support Vector Machine (SVM), Random Forest (RF) and Bayesian Network (BN) for detection and diagnosis of breast cancer. The Original Wisconsin Breast Cancer was used as a dataset and Weka software was used as a Machine Learning tool. The key performance parameters of machine learning classifiers have been compared according to accuracy, recall, precision and ROC area. They have suggested that BN has the best performance according to recall and precision values and RF technique has optimum performance in term of ROC area [5].

Ahmad et al. have exercised machine learning algorithms for predicting the rate of two years recurrence of breast cancer disease. The dataset has been obtained from Iranian Center of Breast Cancer (ICBC) program, collected the time period of 1997-2008 years. The dataset is consisted of population characteristics and 22 input variables also the cases have been collected from 1189 women of diagnosed breast cancer. Artificial Neural Network (ANN), Support Vector Machine (SVM) and Decision Tree (DT) have been applied and SVM has been showed the best performance with highest accuracy and least error rate [6].

Bektas and Babur have studied on diagnosis of breast cancer using machine learning techniques. Kent Ridge Microarray has been used 2 datasets and support vector machine, k-star, random forest algorithm and voted perceptron have been applied. Random forest algorithm has been showed more performance than applied feature selection method [7].

Chen et al. have applied Support Vector Machine classification algorithm on Wisconsin Diagnostic Breast Cancer dataset. In the study, the training and testing sets have been split as 50-50%, 70-30% and 80-20%. According to different training/testing percent, accuracy values have been calculated [8].

In this paper, as SVM and ANN two of the most popular machine learning techniques are applied on Wisconsin Breast Cancer (Original) dataset and the result of applied machine learning (ML) techniques are compared according to performance metrics. The rest of the paper is arranged as follows: Section II explains used material. Section III describes the fundamental points of machine learning techniques. The results are explained in Section IV. The paper ends in conclusion by proposing concluding remarks.

II. DATA SET

In this study, Wisconsin Breast Cancer (Original) (WBC) dataset is acquired from UCI Machine Learning Repository and analyzed. The used dataset is consisting of 699 instances

978-1-7281-1013-4/19/$31.00 ©2019 IEEE

that were classified as benign and malignant. Also, dataset has 11 integer-valued attributes [9].

The common features of dataset are;

- Benign class: 458 (65.5 %)
- Malignant class: 241 (25.5%)

The 11 integer-valued attributes of dataset are;

- Sample code number
- Clump Thickness (1-10)
- Cell Size Uniformity (1-10)
- Cell Shape Uniformity (1-10)
- Marginal Adhesion (1-10)
- Single Epithelial Cell Size (1-10)
- Bare Nuclei (1-10)
- Bland Chromatin (1-10)
- Normal Nuclei (1-10)
- Mitoses (1-10)
- Class (2: Benign, 4: Malignant)

III. EXPERIMENTAL ANALYSIS

In this paper, we have applied SVM and ANN techniques for prediction of the classification of breast cancer to find which machine learning methods performance is better.

Support Vector Machines (SVMs) have been first explained by Vladimir Vapnik and the good performances of SVMs have been noticed in many pattern recognition problems. SVMs can indicate better classification performance when it is compared with many other classification techniques [10]. SVM is one of the most popular machine learning classification technique that is used for the prognosis and diagnosis of cancer. According to SVM, the classes are separated with hyperplane that is consisted of support vectors that are critical samples from all classes. The hyperplane is a separator that is identified as decision boundary among the two sample clusters. SVM can be used for classifying tumors as benign or malignant based on patient's age and tumors size [11]. Artificial Neural Network (ANN) can be expressed in terms of biological neuron system. Especially, it is similar to human brain process system. It is consisted of a lot of nodes that connect each node [12]. ANN have the ability of modelling typical and powerful non-linear functions. It is consisted of a network of lots of artificial neurons. Each of these combinations are comprised of input/output characteristics that perform a local mathematical function. The function could be a computation of weighted sums of inputs which generates an output if it goes beyond a given threshold value. The output could be an input to other neurons in the network. This transaction iterates until the latest output is produced [13]. The authors also published several comparative results in this area [13, 14]. As machine learning techniques, SVM and ANN are applied with WEKA machine learning tool. WEKA is Java based and open source tool. It provides many machine learning algorithms and methods for analysis. It contains many machine learning tools for classification, clustering, regression, association rules mining and visualization. In this study, ARFF (Attribute-Relation File Format) was used for classification of breast cancer. SMO (Sequential Minimal Optimization) algorithm and LibSVM are used as the classification of SVM in Weka software. Also, Multi-Layer Perceptron and Voted Perceptron are used as ANN classifier in Weka.

The results of the applied machine learning techniques application on WBC dataset are reported. We applied k=10-fold cross validation and percentage split (% 66 and % 33 splits) which are training options that split the dataset into a training set to train model and a testing set to evaluate it. In WEKA software, SVM and ANN techniques are applied on breast cancer dataset. The experimental results are demonstrated in Table 1 and Table 2 respectively. In the results, performance metrics of accuracy, precision, recall and ROC Area are compared.

Table 1: The experimental results of Support Vector Machine classification techniques.

Support Vector Machine	**The results of performance**				
	Test Options	***Accuracy***	***Precision***	***Recall***	***ROC Area***
Sequential Minimal Optimization (SMO)	k=10 cross validation	0,969957	0,97	0,97	0,968
	Percentage split	0,953782	0,954	0,954	0,949
LibSVM	k=10 cross validation	0,957082	0,96	0,957	0,964
	Percentage split	0,957983	0,958	0,958	0,957

Table 2: The experimental results of Artificial Neural Network classification techniques.

Artificial Neural Network	**The result of performance**				
	Test Options	***Accuracy***	***Precision***	***Recall***	***ROC Area***
Multi-Layer Perceptron (MLP)	k=10 cross validation	0,95442	0,954	0,954	0,988
	Percentage split	0,953782	0,955	0,954	0,994
Voted Perceptron	k=10 cross validation	0,909871	0,919	0,91	0,929
	Percentage split	0,882353	0,899	0,882	0,915

IV. CONCLUSION

Breast Cancer is the most frequent disease as a cancer type for women. Therefore, any development for diagnosis and prediction of cancer disease is capital important for a healthy life. In this paper, we have discussed two popular machine learning techniques for Wisconsin Breast Cancer classification. Artificial Neural Network and Support Vector Machine are used as ML techniques for the classification of WBC (Original) dataset in WEKA tool. The effectiveness of applied ML techniques is compared in term of key performance metrics such as accuracy, precision, recall and ROC area. Based on the performance metrics of the applied ML techniques, SVM (Sequential Minimal Optimization Algorithm) has showed the best performance in the accuracy of 96,9957 % for the diagnosis and prediction from WBC dataset.

REFERENCES

[1] Cancer, https://www.who.int/en/news-room/fact-sheets/detail/cancer. Last Access: 25.01.2019.

[2] Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, *Ca-a Cancer Journal for Clinicians*, *68* (1), pp. 7-30.

[3] Maity, N. G., & Das, S. (2017). Machine learning for improved diagnosis and prognosis in healthcare. In *2017 IEEE Aerospace Conference*, pp. 1-9.

[4] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, *83*, pp. 1064-1069.

[5] Bazazeh, D., & Shubair, R. (2016). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In *2016 5th International Conference on Electronic Devices, Systems and Applications,* pp. 1-4.

[6] Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013). Using three machine learning techniques for predicting breast cancer recurrence. *J Health Med Inform*, *4* (124).

[7] Bektas, B., & Babur, S. (2016). Machine learning based performance development for diagnosis of breast cancer, *Medical Technologies National Congress,* pp. 1-4.

[8] Chen, H. L., Yang, B., Liu, J., & Liu, D. Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications*, *38* (7), pp. 9014-9022.

[9] UCI Breast Cancer Wisconsin (Original) Dataset, https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29. Last Access: 30.01.2019.

[10] Huang, M. W., Chen, C. W., Lin, W. C., Ke, S. W., & Tsai, C. F. (2017). SVM and SVM ensembles in breast cancer prediction. *PloS one*, *12* (1).

[11] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction, *Computational and structural biotechnology Journal*, *13*, pp. 8-17.

[12] Umadevi, S., & Marseline, K. J. (2017). A survey on data mining classification algorithms. In *2017 International Conference on Signal Processing and Communication,* pp. 264-268.

[13] Padmapriya S., Devika M., Meena V., Dheebikaa S. B., & Vinodhini R. (2016). Survey on Breast Cancer Detection Using Weka Tool. *Imperial Journal of Interdisciplinary Research (IJIR)*, Vol 2, no. 4.

[14] S. Turgut, M. Dagtekin, T. Ensari, Microarray Breast Cancer Data Classification Using Machine Learning Methods, Int. Conf. on Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, DOI: 10.1109/EBBT.2018.8391468, April 18-19, 2018.

[15] M. Amrane, S. Oukid, I. Gagaoua, T. Ensari, Breast Cancer Classification Using Machine Learning, Int. Conf. on Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, DOI: 10.1109/EBBT.2018.8391453, April 18-19, 2018.