Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 7, S. 91-95, 3, 2023
© Telif hakkı IJANSER'e aittir
Araştırma Makalesi

https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 7, pp. 91-95, 3, 2023
Copyright © 2023 IJANSER
Research Article

Breast Cancer Detection Using Machine Learning Algorithms

Ahmet Enes KILIÇ [1], Murat KARAKOYUN [2]

[1]*Computer Engineering, Necmettin Erbakan University, KONYA*
[2]*Computer Engineering, Necmettin Erbakan University, KONYA*

[1]*(aekilica@gmail.com) Email of the corresponding author*

(Received: 25 March 2023, Accepted: 12 April 2023)

(2nd International Conference on Scientific and Academic Research ICSAR 2023, March 14-16, 2023)

ATIF/REFERENCE: Kılıç, A. E. & Karakoyun, M. (2023). Breast Cancer Detection Using Machine Learning Algorithms. *International Journal of Advanced Natural Sciences and Engineering Researches*, 7(3), 91-95.

***Abstract* –** Breast cancer is the most common type of cancer among women worldwide and has the highest mortality rate among women. As early diagnosis is important in cancer, early diagnosis in breast cancer significantly reduces the death rate. Thus, early detection of breast cancer significantly increases the chances of survival. Early diagnosis of breast cancer can significantly increase the chances of survival, as it can encourage timely clinical treatment. In this study, the data quality of the Breast Cancer Wisconsin (Diagnostic) dataset, which includes metric data extracted from the biopsy piece with various data mining methods was increased and the patient's breast cancer was classified as benign or malignant with machine learning algorithms. When we compare the developed machine learning algorithms; K-Nearest Neighbor algorithm showed higher performance than other machine learning algorithms with 99.3% accuracy, 98.9% precision, 100% recall and 99.4% f1-score values. The second most successful model on the test set is Support Vector Machine and Logistic Regression.

Keywords – Breast Cancer, Machine Learning, Data Mining, Breast Cancer Wisconsin Dataset

I. INTRODUCTION

Breast cancer has been identified as largest cause of cancer deaths among middle-aged women. According to the projection of the World Health Organization, the estimated number of breast cancer diagnoses among women is 1.5 million each year, with 500,000 women dying from breast cancer in 2015 [1]. Early detection of breast cancer is important in order to reduce the mortality rates due to breast cancer in women. There are many early detection strategies, such as screening, to detect breast cancer early. In addition, with the development of artificial intelligence, various machine learning techniques have been developed. With these techniques, the decisions of experts in most fields can be supported. The use of machine learning techniques is increasing rapidly, helping medical professionals diagnose disease [2]. In breast cancer research, machine learning algorithms can be used to detect and predict cancer.

In this study is aimed to classify the patient's breast cancer as benign or malignant with machine learning algorithms by increasing the data quality of the Breast Cancer Wisconsin (Diagnostic) data set, which includes metric data extracted from the biopsy piece by various data mining methods.

The rest of the study is as follows: In Section 2, previous studies similar to this study are mentioned, in Section 3 the details of the dataset used and the machine learning methods used in the study, in Section 4 the experimental results obtained as a result of the study, and finally the conclusion part.

In this section, some studies in the literature are mentioned. The machine learning algorithms and results in the literature developed using the Breast Cancer Wisconsin dataset are shown in Table 1. According to Table 1, generally Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbour (K-NN) and Naïve Bayes (NB) machine learning algorithms were used. When we look at the general evaluation, it has been seen that the performance of the algorithms in the literature varies in general due to the changes in data preparation or data pre-processing steps.

Table 1. Some studies in the literature

Studies	**Year**	**Algorithm and Performance**
[3]	2022	SVM: 0.95 K-NN: 0.94 NB: 0.95 RF: 0.96 Logistic Regression (LR): 0.95 DT: 0.94 ANN: 0.95
[4]	2022	PCA-SVM: 0.98 LDA-SVM: 0.98 RF-LDA: 0.95 RF-PCA: 0.97
[5]	2021	SVM: 0.97 RF: 0.96 LR: 0.96 DT: 0.95 K-NN: 0.94
[6]	2020	LR: 0.98 K-NN: 0.96
[7]	2020	DT: 0.99 NB: 0.99
[8]	2020	RF: 0.99 Gradient Boosted Trees: 0.96 LR: 0.98 DT: 0.95 SVM: 0.93
[9]	2019	SVM: 0.96 Multilayer Perceptron (MP):0.95 Voted Perceptron: 0.91
[10]	2018	K-NN: 0.99 SVM: 0.91 LR: 0.90
[11]	2018	K-NN: 0.95 SVM: 0.98 DT: 0.93
[12]	2018	Review Article
[13]	2018	DT: 0.93 NB: 0.97 RBF Network: 0.97
[14]	2018	SVM: 0.93 MP: 0.98 K-NN: 0.91 DT: 0.97
[15]	2017	SVM: 0.98 K-NN: 0.97
[16]	2017	SVM: 0.99 K-NN: 0.96 NB: 0.93 RF: 0.98 LR: 0.95
[17]	2017	DT: 0.94
[18]	2016	K-Means: 0.92
[19]	2016	SVM: 0.97 K-NN: 0.95 NB: 0.96
[20]	2016	NB: 0.92 Neural Networks: 1.00 SVM+DT: 0.95 Fuzzy: 0.93 RelevanceVectorMachine: 0.97
[21]	2015	SVM: 0.96 MP: 0.95
[22]	2015	DT: 0.96 Bayesian Networks: 0.97
[23]	2014	DT: 0.94
[24]	2012	MP: 0.95 K-NN: 0.94 DT: 0.95 NB: 0.96

II. MATERIALS AND METHOD

A. *Dataset*

There are 699 samples in the Breast Cancer Wisconsin dataset from UCI repository [25], and each sample has 10 features and 1 class information. In addition, there are 16 missing feature values specified as '?' in the dataset. Dataset description is also included in Table 2.

Table 2. Dataset description

Features	**Value Range**
Sample code number	id number
Clump Thickness	1 - 10
Uniformity of Cell Size	1 - 10
Uniformity of Cell Shape	1 - 10
Marginal Adhesion	1 - 10
Single Epithelial Cell Size	1 - 10
Bare Nuclei	1 - 10
Bland Chromatin	1 - 10
Normal Nucleoli	1 - 10

Mitoses	1 - 10
Class	2: benign 458 (65.5%) 4: malignant 241 (34.5%)

B. *Method*

The value range of the properties of the dataset is in the range of 1 to 10. The missing data in various properties of 16 samples were filled with an average value of 5 using the missing data filling method, which is a pre-processing method commonly used in data mining. Input and output values were created by separating the feature and class values in the dataset. 80% of the generated input and output values are divided as training (599 samples) and 20% as test data (140 samples). The prepared data is tested by various machine learning algorithms such as K-NN [26], DT [27], SVM [28], NB [29], RF [30], LR [31], MP [32], and the results are shown in section 5. Figure 1 shows the algorithmic representation of the study.

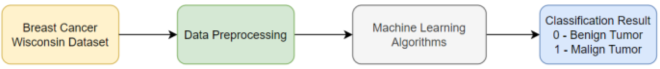

Fig. 1 Algorithmic representation of the study

C. *Evaluation Metrics*

The Breast Cancer Wisconsin dataset, which was prepared using data mining methods, was tested with various machine learning algorithms. In this study, accuracy, precision, recall and f1 score metrics, which are calculated in Eqs. (1), (2), (3) and (4), respectively, are considered as evaluation criteria. The equations given below show the metric calculations according to the confusion matrix extracted in Table 3.

Table 3. Confusion matrix

		Actual Values	
		Benign	Malign
Predicted Values	Benign	TP	FN
	Malign	FP	TN

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (1)$$

$$Precision = \frac{TP}{TP+FP} \qquad (2)$$

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

$$f1-score = 2 * \frac{precision*recall}{precision+recall} \qquad (4)$$

III. RESULTS

The results of the machine learning models trained to classify benign and malignant breast cancer on the Breast Cancer Wisconsin dataset are given in Table 4. According to Table 4, while the K-NN machine learning algorithm gave the highest classification performance for benign and malignant breast cancer, the DT machine learning algorithm showed the lowest classification performance. In addition, the confusion matrix extracted for each algorithm is shown in Figure 2 to Figure 8. In the confusion matrices below, 0 represents benign tumour and 1 represents malignant tumour.

Table 4. Machine Learning Algorithm Results

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-S.**
K-NN	**0.993**	**0.989**	**1.000**	**0.994**
DT	0.957	0.947	0.989	0.967
SVM	0.986	**0.989**	0.989	0.989
NB	0.971	**0.989**	0.967	0.978
RF	0.978	0.978	0.989	0.983
LR	0.986	0.978	**1.000**	0.989
MP	0.978	0.968	**1.000**	0.984

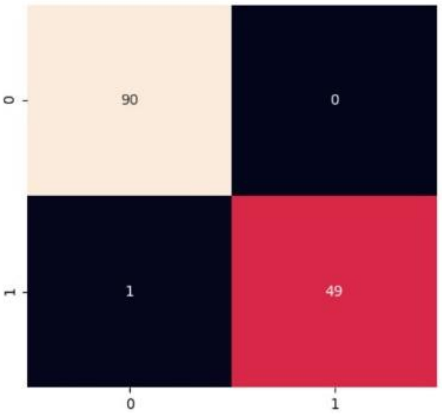

Fig. 2 Confusion Matrix of K-NN

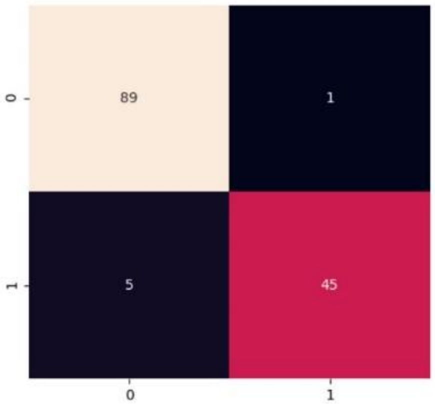

Fig. 3 Confusion Matrix of DT

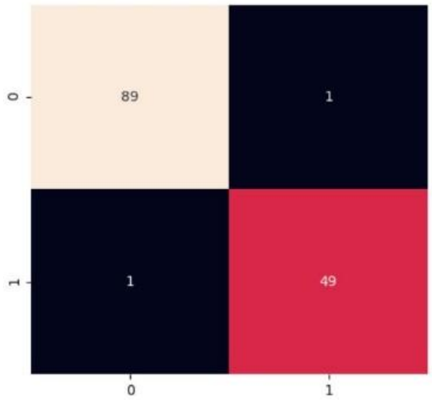

Fig. 4 Confusion Matrix of SVM

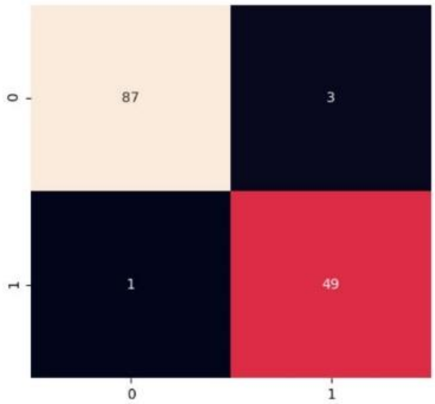

Fig. 5 Confusion Matrix of NB

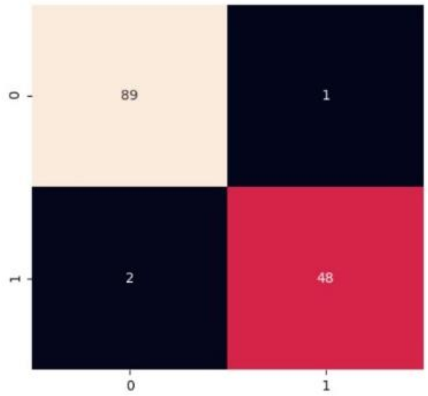

Fig. 6 Confusion Matrix of RF

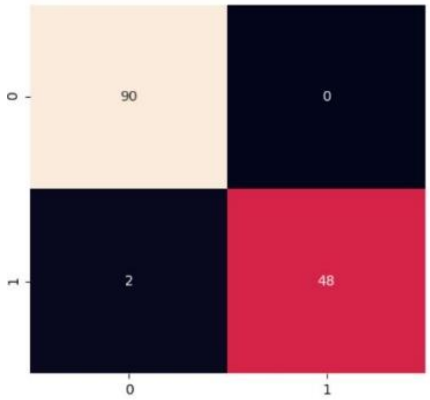

Fig. 7 Confusion Matrix of LR

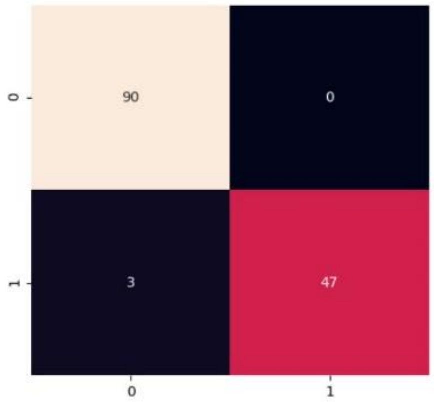

Fig. 8 Confusion Matrix of MP

IV. CONCLUSION

It has been revealed by the literature research that machine learning is widely used in the field of medicine as in many different fields and is used as a decision support system in the diagnosis of diseases. Its use is increasing, especially in the diagnosis of cancer. Breast cancer is the most common type of cancer among women and poses a risk of death if not detected early. For this reason, as can be seen from the studies in the literature, it is important to detect the diagnosis of breast cancer accurately and with high performance. Studies with this data set in the literature were examined and a comparison of different accuracies between machine learning algorithms was given. It has been observed that the reason for this is that the differences in the preparation of the data or the pre-processing of the data affect the results. In this study, the breast cancer of the patient was classified as benign/malignant using the Breast Cancer Wisconsin dataset and data mining and machine learning algorithms, and a comparison was made by looking at certain metrics in various machine learning algorithms. In future studies, multiple classifications are planned on a more comprehensive data set.

REFERENCES

[1] Bataineh, A. Al. "A comparative analysis of nonlinear machine learning algorithms for breast cancer detection." *International Journal of Machine Learning and Computing* 9.3 (2019): 248-254.

[2] Sevli, O. "Performance Comparison of Different Machine Learning Techniquesin Diagnosis of Breast Cancer." *Eur. J. Sci. Technol.* 16 (2019): 176-185.

[3] El Massari, Hakim, et al. "An ontological model based on machine learning for predicting breast cancer."

International Journal of Advanced Computer Science and Applications 13.7 (2022).

[4] Egwom, Onyinyechi Jessica, et al. "An LDA–SVM Machine Learning Model for Breast Cancer Classification." *BioMedInformatics* 2.3 (2022): 345-358.

[5] Naji, Mohammed Amine, et al. "Machine learning algorithms for breast cancer prediction and diagnosis." *Procedia Computer Science* 191 (2021): 487-492.

[6] JIJITHA, S.; AMUDHA, Thangavel. "Breast cancer prognosis using machine learning techniques and genetic algorithm: experiment on six different datasets." In: *Evolutionary Computing and Mobile Sustainable Networks: Proceedings of ICECMSN 2020*. Springer Singapore, 2021. p. 703-711.

[7] MOHAMMED, Siham A., et al. "Analysis of breast cancer detection using different machine learning techniques. " In: *Data Mining and Big Data: 5th International Conference, DMBD 2020*, Belgrade, Serbia, July 14–20, 2020, Proceedings 5. Springer Singapore, 2020. p. 108-117.

[8] Rasool, Mir Junaid, Amanpreet Singh Brar, and Hardeep Singh Kang. "Risk prediction of breast cancer from real time streaming health data using machine learning." *Int. Res. J. Mod. Eng. Technol. Sci* 2 (2020): 409-418.

[9] BAYRAK, Ebru Aydındag; KIRCI, Pınar; ENSARI, Tolga. "Comparison of machine learning methods for breast cancer diagnosis." In: *2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT). IEEE*, 2019. p. 1-3.

[10] KUMARI, Madhu; SINGH, Vijendra. "Breast cancer prediction system." *Procedia computer science*, 2018, 132: 371-376.

[11] Obaid, Omar Ibrahim, et al. "Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer." (2018).

[12] Yue, Wenbin, et al. "Machine learning with applications in breast cancer diagnosis and prognosis." *Designs* 2.2 (2018): 13.

[13] Chaurasia, Vikas, Saurabh Pal, and B. B. Tiwari. "Prediction of benign and malignant breast cancer using data mining techniques." *Journal of Algorithms & Computational Technology* 12.2 (2018): 119-126.

[14] GUPTA, Madhuri; GUPTA, Bharat. "A comparative study of breast cancer diagnosis using supervised machine learning techniques." In: *2018 second international conference on computing methodologies and communication (ICCMC). IEEE*, 2018. p. 997-1002.

[15] ISLAM, Md Milon, et al. "Prediction of breast cancer using support vector machine and K-Nearest neighbors." In: *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC). IEEE*, 2017. p. 226-229.

[16] SHAHNAZ, Celia, et al. "Efficient approaches for accuracy improvement of breast cancer classification using wisconsin database." In: *2017 IEEE region 10 humanitarian technology conference (R10-HTC). IEEE*, 2017. p. 792-797.

[17] YI, Liu; YI, Wu. "Decision tree model in the diagnosis of breast cancer." In: *2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC). IEEE*, 2017. p. 176-179.

[18] Dubey, Ashutosh Kumar, Umesh Gupta, and Sonal Jain. "Analysis of k-means clustering approach on the breast cancer Wisconsin dataset." *International journal of computer assisted radiology and surgery* 11 (2016): 2033-2047.

[19] Asri, Hiba, et al. "Using machine learning algorithms for breast cancer risk prediction and diagnosis." *Procedia Computer Science* 83 (2016): 1064-1069.

[20] GAYATHRI, B. M.; SUMATHI, C. P. "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer." In: *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE*, 2016. p. 1-5.

[21] GHOSH, Soumadip; MONDAL, Sujoy; GHOSH, Bhaskar. "A comparative study of breast cancer detection based on SVM and MLP BPN classifier." In: *2014 First International Conference on Automation, Control, Energy and Systems (ACES). IEEE*, 2014. p. 1-4.

[22] Rodrigues, L., “Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection”, Proceedings of XI Workshop de Visão Computacional, 2015.

[23] Sumbaly, Ronak, N. Vishnusri, and S. Jeyalatha. "Diagnosis of breast cancer using decision tree data mining technique." *International Journal of Computer Applications* 98.10 (2014).

[24] SALAMA, Gouda I.; ABDELHALIM, M. B.; ZEID, Magdy Abd-elghany. "Experimental comparison of classifiers for breast cancer diagnosis." In: *2012 Seventh International Conference on Computer Engineering & Systems (ICCES). IEEE*, 2012. p. 180-185.

[25] (2007) The IEEE website. UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

[26] Peterson, Leif E. "K-nearest neighbor." *Scholarpedia* 4.2 (2009): 1883.

[27] Kingsford, Carl, and Steven L. Salzberg. "What are decision trees?." *Nature biotechnology* 26.9 (2008): 1011-1013.

[28] Hearst, Marti A., et al. "Support vector machines." *IEEE Intelligent Systems and their applications* 13.4 (1998): 18-28.

[29] KAMEL, Hajer; ABDULAH, Dhahir; AL-TUWAIJARI, Jamal M. "Cancer classification using gaussian naive bayes algorithm." In: *2019 International Engineering Conference (IEC). IEEE*, 2019. p. 165-170.

[30] Boulesteix, Anne-Laure, et al. "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.6 (2012): 493-507.

[31] Steyerberg, Ewout W., et al. "Internal validation of predictive models: efficiency of some procedures for logistic regression analysis." *Journal of clinical epidemiology* 54.8 (2001): 774-781.

[32] Murtagh, Fionn. "Multilayer perceptrons for classification and regression." *Neurocomputing* 2.5-6 (1991): 183-197.