Advances in Artificial Intelligence Research (AAIR)
ISSN 2757-7422
www.dergipark.com/aair/

Vol. 2 (No. 2), pp. 65-70, 2022
doi: 10.54569/aair.1142519
Published online: Sept 15, 2022

Using Classification Algorithms in Data Mining in Diagnosing Breast Cancer

Büşranur Nalbant [1,*], İrem Düzdar Argun [2]

[1] Düzce University, Department of Mechatronics Engineering, Düzce, Turkey
[2] Düzce University, Department of Industrial Engineering, Düzce, Turkey

Abstract

Data mining is the process of extracting useful information from large-scale data in an understandable and logical way. The main machine learning techniques of data mining are classification and regression, association rules and cluster analysis. Classification and regression are known as predictive models; clustering and association rules are known as descriptive models. In this study, the classification method is used. With this method, it is aimed to assign a data set to one of the previously determined different classes. The data set used in this study is obtained from the UCIrvine Machine Learning Repository database. The dataset named as “Breast cancer” consists of 699 samples and 10 features collected by William H. at the University of Wisconsin Hospital. The dataset content includes information about the characteristics of some cells analyzed in the detection of breast cancer. The goal of this study is to make a classification by determining whether one has cancerous or non-cancerous cells. In this study, data mining analyzes are performed in WEKA and Orange programs using SVM (Support Vector Machine) and Random Forest algorithms. According to the analysis results, a comparison is made on the data set regarding the previous studies. It is aimed that the conclusions obtained at the end of the study will guide medical professionals working in the diagnosis of breast cancer.

Keywords: *Data Mining; classification algorithms; breast cancer.*

1. Introduction

Cancer is a general term used for all diseases that occur when cells in an organ or tissue in the human body begin to multiply uncontrollably. According to the researches, among the common cancer types, breast cancer is the second most common cancer type worldwide after lung cancer [1]. Considering the 2020 data of the World Health Organization, International Agency for Research on Cancer (IARC), 1 out of every 8 cancer types reported as breast cancer in 2020; 2.3 million breast cancer cases were diagnosed and 685.000 people died. Moreover; this type of cancer was in the 5th place in the world among the other types of cancer [2].

Correct diagnosis of diseases consists of a complex process. Medical professionals use biochemical tests and radiology to make a diagnosis. These methods may vary according to the diseases. Breast ultrasound is of great importance in the diagnosis of breast cancer. It is a preferred cancer prediction method because it is painless and it does not contain radiation [3]. This method, which is widely used, is performed with computer aided diagnostic tools. Thanks to these computer-assisted diagnostic tools, it is determined whether the mass in the patient is benign or malignant [4].

In this study, it is aimed to diagnose the disease by determining whether the mass in the patient is benign or malignant with the classification process - which is one of the machine learning methods. The objective of this study is to provide benefits to the experts by minimizing the loss of time before exceeding the vital stage. Because early diagnosis of breast cancer is of great importance so as to get positive treatment results.

2. Literature Review

As a result of the literature review, it has been revealed that many studies have been carried out on the Breast Cancer dataset since 2004.

Law vd. (2004) [5] suggested the use of mixture-based clustering algorithm in his study and tested it on the data set. The classification accuracy of the algorithm was 90.7%. Luukka and Leppälampi [2006) [6] used the C4.5 classification algorithm for breast cancer diagnosis. It reached a success value of 94.06%. Li and Lu (2010) [7] first used principal component analysis (PCA) to reduce feature sizes in the data set and then proposed a class probability-based kernel (CPBK) method based on Support Vector Machines (SVM). It reached an accuracy value of 93.26%. Lavanya and Rani (2011) [8] used the classification and regression tree (CART) algorithm to achieve the best success in the data set with a value of 94.84%. In the same year, Maldonado vd. [9] used a recursive dimension elimination (SVM-RFE) based technique. The average

*Corresponding author
E-mail address: busranur73390@ogr.duzce.edu.tr

Received: 08/Jul/2022; Received in revised form: 08/Sept/2022; Accepted: 14/Sept/2022.

classification accuracy of this method was 95.25%.

Considering the historical development of data mining and developing technology, these studies are exemplary. However; when the studies conducted in recent years are examined, it is seen that the early diagnosis of the disease has more increased with the developing technology. Therefore; recent studies promise great hopes.

Takci (2016) [10] conducted a study with three separate data sets, including the Wisconsin data set. He made various comparisons between machine learning methods and Centroid Classifiers. He also reported the results in terms of accuracy and time. Euclidean-based center classifier gave the highest classification accuracy with a value of 99.04%. Akyol (2018) [11] investigated the importance of features using the Recursive Feature Selection method on the data set and used Random Forest and Logistic Regression classifier algorithms. The learning process, which consisted of testing and training stages, was carried out by using the 5-fold cross-validation technique. As a result of the study, it was shown that the best classification success (98% accuracy) was obtained with the Random Forest algorithm.

Karaci (2019) [12] developed a DNN model (deep neural network) for breast cancer diagnosis using some data such as body mass index, insulin and age glucose. Data were obtained from 116 women, 52 healthy and 64 with breast cancer. Then; machine learning was carried out with the obtained data. This model estimated healthy women as a minimum of 88.2% and a maximum of 94.1%. It also estimated women with breast cancer as a minimum of 88.8% and a maximum of 94.4%. In the study conducted by Kor (2019) [13], it was determined that the SVM method had the highest rate of classifying breast cancer as benign and malignant with 97.66%. Yavuz and Eyuboglu (2019) [14] proposed a score fusion method based on Generalized Regression Neural Network (GRNN) and Feed Forward Neural Network (FFNN) to classify breast cancer data samples as benign or malignant. The usefulness of these two main nets and the proposed method were examined; the performance results were presented comparatively. In another study conducted in the same year, Sevli (2019) [15] created confusion matrices and ROC curves after the training process with various machine learning methods and then compared the success of each technique. As a result of this comparison, it was revealed that logistic regression was the most successful method with an accuracy rate of 98.24%.

Cengil and Cinar (2020) [16] used Keras Deep Learning Library tools for classification process. The application results showed that the classification performance was around 98%. Akcan and Sertbas (2021) [17] used these machine learning methods: Support Vector Machine (SVM), K-Nearest Neighborhood (KNN), Naive Bayes (NB), Decision Tree (DT), Adaboost (SVC), XGBoost and Random Forest (RF). Among them, Adaboost (SVC) and XGBoost had the highest success rate with the same accuracy of 97.37%.

As a result of the literature research, it can be seen that machine learning methods has been widely used in the field of medicine. Therefore; there have been many publications about research on the diagnosis of breast cancer. In this study, like previous studies, it is hoped to be a promising study for medical professionals in the diagnosis of medical disease.

3. Material and Method

3.1. Materials

In this study, breast cancer data collected by William H. at the University of Wisconsin Hospital is used. The dataset is obtained from the UCIrvine Machine Learning Repository database. Both two different software, WEKA and Orange software, and two different data mining algorithms, Support Vector Machine (SVM) and Random Forest (RF) algorithm, are used.

On the data set, models are created with the algorithms of the classification methods specified by a computer with an Intel Core i7 processor and 12 GB RAM. Then; machine learning is carried out by using programming languages. The performance rates obtained from the algorithms are compared by considering the results of the previous studies mentioned in the literature review. At last; the performance results of the algorithms and software on the data set are evaluated.

3.1.1. Data set

The data set includes 699 samples. The number of attributes is 10. It does not contain any qualifications with incomplete information. The class distribution of these 699 data is 458 samples as benign and 241 samples as malignant. The data are obtained by digitizing the images of the mass seen in the chest. In **Table 1**, value ranges, means and standard deviation values of 10 features (closure thickness, size uniformity, shape uniformity, adhesion, epithelial size, bare nucleus, soft chromatin, normal nucleoli, mitosis and class) in the given data set are shown.

Table 1. *Attribute Descriptions and Values*

Attribute Description	Value	Mean	Standard Deviation
Closing Thickness	1-10	4.442	2.820
Dimensional Isomorphism	1-10	3.150	3.065
Figurative Isomorphism	1-10	2.840	2.988
Adhesion	1-10	3.234	2.864
Epithelial Dimension	1-10	3.544	2.223
Naked Nucleus	1-10	3.445	3.449
Soft Chromatin	1-10	2.869	3.050
Normal Nucleoli	1-10	2.869	3.050
Mitosis	1-10	1.603	1.732
Class	2-4		

3.2. Classifiers

The algorithms used in this study are among the popular methods in data mining. Success results are taken into account in the selection of these algorithms. There are many studies in the literature proving Support Vector Machines (SVM) success. The SVM algorithm is a classification algorithm used to separate data which belongs to two separate classes accordingly with each other [18]. The Random Forest (RF) algorithm, the second algorithm used in the study, is also a widely used method in the classification process. It is an ensemble learning algorithm that creates many decision trees and determines the most suitable one [19]. There is information about the support vector machine algorithm and random forest algorithm used in this study below.

3.2.1. Support Vector Machine Algorithm

Created by Vladimir Vapnik, Support Vector Machines (SVM) is a new forward routing network. The SVM's powerful tools used to solve many common problems and drive many current developments for the detailed kernel are its uncomputed, predictive and low rate function. Statistical education and treatment risk are minimized [20].

In two dimensional space linear separation mechanisms, in three dimensional space planar separation and in multidimensional separation in hyperplane, the data can be grouped in more than one group by SVM. The case where the data group can be separated by a line is when the group can be separated linearly. The idea here is that the object separating the two classes is a corridor rather than a line; furthermore, corridor's width is determined by some data vector to be the largest possible width [21].

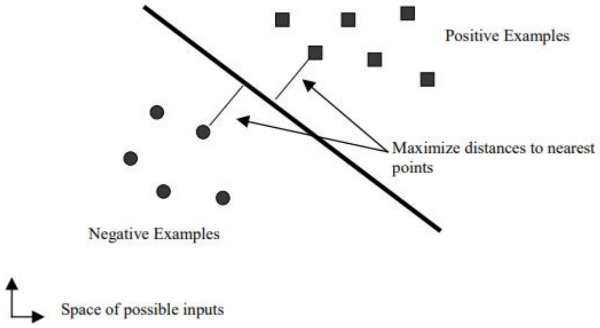

Figure 1. *Support Vector Machines (SVM) [22]*

3.2.2. Random Forest Algorıthm

The Random Forest algorithm was developed by Breiman in 2001. RF is a classification model that tries to make more accurate classification by producing more compatible models using multiple decision trees. By bringing together, these formed decision trees constitute the decision forest. The created decision trees are randomly determined subsets of the dataset in relation. It offers excellent validity. It has more precise results than Adaboost and Support Vector Machines for many datasets [23]. It works at four steps;

- Random samples are selected from a given dataset.

- A decision tree is created for each sample and a prediction result is taken from each decision tree.
- A vote is taken for each predicted outcome.
- The result of the prediction is chosen by using the most votes as the final prediction.

In **Figure 2**, the tree structure is shown according to the results obtained from the RF algorithm in the Orange application of the data set used in the study.

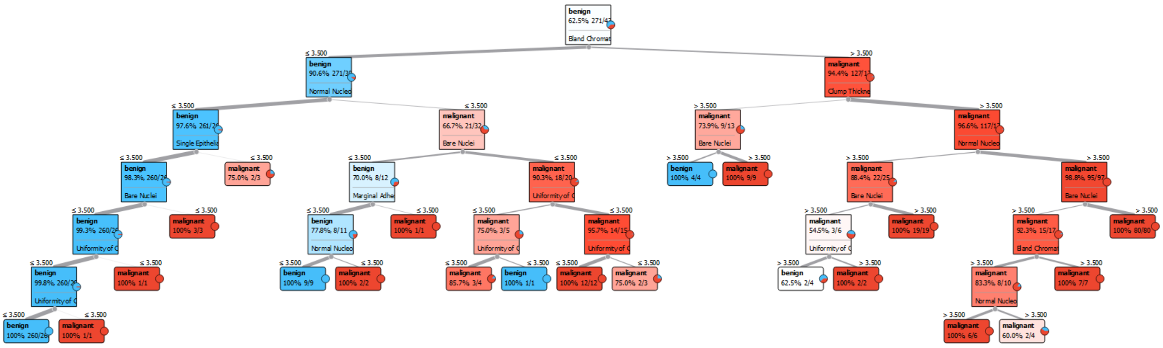
benign 62.5% 271/43 Bland Chromat
≤ 3.500 benign 90.6% 271/30 Normal Nucleo
> 3.500 malignant 94.4% 127/1 Clump Thickne
≤ 3.500 benign 97.6% 261/2 Single Epitheli
≤ 3.500 malignant 66.7% 21/32 Bare Nuclei
> 3.500 malignant 73.9% 9/13 Bare Nuclei
> 3.500 malignant 96.6% 117/1 Normal Nucleo
≤ 3.500 benign 98.3% 260/2 Bare Nuclei
≤ 3.500 malignant 75.0% 2/3
≤ 3.500 benign 70.0% 8/12 Marginal Adhe
≤ 3.500 malignant 90.3% 18/20 Uniformity of C
> 3.500 benign 100% 4/4
> 3.500 malignant 100% 9/9
> 3.500 malignant 88.4% 22/25 Bare Nuclei
> 3.500 malignant 98.8% 95/97 Bare Nuclei
≤ 3.500 benign 99.3% 260/2 Uniformity of C
≤ 3.500 malignant 100% 3/3
≤ 3.500 benign 77.8% 8/11 Normal Nucleo
≤ 3.500 malignant 100% 1/1
≤ 3.500 malignant 75.0% 3/5 Uniformity of C
≤ 3.500 malignant 95.7% 14/15 Uniformity of C
> 3.500 malignant 54.5% 3/6 Uniformity of C
> 3.500 malignant 100% 19/19
> 3.500 malignant 92.3% 15/17 Bland Chromat
> 3.500 malignant 100% 80/80
≤ 3.500 benign 99.8% 260/2 Uniformity of C
≤ 3.500 malignant 100% 1/1
≤ 3.500 benign 100% 9/9
≤ 3.500 malignant 100% 2/2
≤ 3.500 malignant 85.7% 3/4
≤ 3.500 benign 100% 1/1
≤ 3.500 malignant 100% 12/12
≤ 3.500 malignant 75.0% 2/3
> 3.500 benign 62.5% 2/4
> 3.500 malignant 100% 2/2
> 3.500 malignant 83.3% 8/10 Normal Nucleo
> 3.500 malignant 100% 7/7
≤ 3.500 benign 100% 260/26
≤ 3.500 malignant 100% 1/1
> 3.500 malignant 100% 6/6
> 3.500 malignant 60.0% 2/4

Figure 2. *Orange Classification tree viewer breast cancer dataset*

4. Results and Discussion

The results of the analysis, which was carried out for the early diagnosis of breast cancer are given in **Table 2**. The performance values (Accuracy, Precision, Recall and F-measure) obtained from the RF algorithm and SVM algorithm of the breast cancer dataset which are analyzed by using Weka and Orange applications are shown in the **Table 2**.

Table 2. *Application algorithm results*

	Method	Accuracv	Precision	Recall	F-Measure
WEKA	RF	96.7096 %	0.980	0.969	0.975
	SVM	95.7082%	0.993	0.941	0.966
Orange	RF	89.9 %	0.959	0.959	0.959
	SVM	87.2%	0.962	0.962	0.962

According to the analysis results in **Table 2**, the accuracy values of all algorithms are above 87%. RF algorithm of Weka software has the highest accuracy value with 96.7096%. It is important to have high accuracy values. Thus; it has observed that it is appropriate to use both algorithms to obtain meaningful information that can be used in the data. In addition to this; the open source Weka program gives higher accuracy values when the software used in this study is compared.

In **Table 2**, the results of the algorithm analysis performed on the data set of the Weka and Orange software used in the study are also given. In **Figure 3**, this table is shown graphically. Looking at the graph, it is seen that the accuracy values are high.

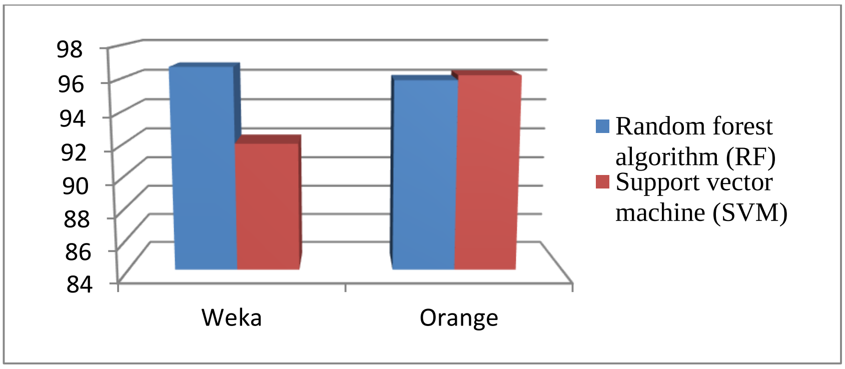

Figure 3. *Results of Accuracy Rates Analysis for Data Mining Techniques*

5. Conclusions

Considering the data of death in cancer, early diagnosis of the disease is vital for the medical field. For this reason; scientific researches play a crucial role. Data mining is, doubtlessly, very helpful for shortening the time during the diagnosis process. The data in the dataset used in this study are obtained by digitizing the images of the mass seen in the chest. In order to get this dataset, two different machine learning algorithms in Weka and Orange software are used. Analysis results are shown in tables and graphics. The software which is used and machine learning algorithms which are applied are compared to each other. According to the comparison result, the highest accuracy value is obtained from the SVM algorithm used in the Weka software.

When the previous studies in the literature review are examined, it is clear that the accuracy values of Adaboost and SVM are generally higher. As a result of this study, it has seen that the values of the SVM algorithm are high. However; RF classifier gives higher results with a success rate of around 94.11% compared to other methods. Therefore; RF is proposed as the most successful method for this dataset. With this study, it is aimed to facilitate the early diagnosis of medical professionals and to minimize the loss of time that may occur during the diagnosis of the disease.

Declaration of interest

It was presented as a summary at the ICAIAME 2022 conference.

Acknowledgements

We thank the UC Irvine Machine Learning Repository database for preparing the breast cancer dataset used in the study.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. *CA Cancer J Clin*. 2018 Nov;68(6):394-424. doi: 10.3322/caac.21492.

[2] Jeleń Ł., Krzyżak A., Fevens T. and Jeleń M., “Influence of feature set reduction on breast cancer malignancy classification of fine needle aspiration biopsies”, *Computers in Biology and Medicine*, 79 (2016) pp. 80-91.

[3] Uzm. Dr. Rengin Türkgüler, [Online]. Available: https://www.drrengin.com/tr/meme-ultranonu (accessed: August 5, 2022).

[4] Mittal S. et al. “Biosensors for breast cancer diagnosis: A review of bioreceptors, biotransducers and signal amplification strategies”, *Biosensors and Bioelectronics* 88 (2017): 217-231.

[5] Law M.H.C., Figueiredo M.A.T. and Jain A.K., “Simultaneous feature selection and clustering using mixture models”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9), (2004) pp. 1154-1166.

[6] Luukka P. and Leppälampi T., “Similarity classifier with generalized mean applied to medical data,” *Computers in Biology and Medicine*, 36(9) (2006), pp. 1026-1040.

[7] Li D.-C. and Liu C.-W., “A class possibility based kernel to increase classification accuracy for small data sets using support vector machines,” *Expert Systems with Applications*, 37(4) (2010), pp. 3104-3110.

[8] Lavanya D. and Rani K.U., “Performance evaluation of decision tree classifiers on medical datasets,” *International Journal of Computer Applications*, 26(4) (2011), pp. 1-4.

[9] Maldonado S., Weber R. and Basak J., “Simultaneous feature selection and classification using kernel-penalized support vector machines”, *Information Sciences*, 181(1) (2011), pp. 115-128.

[10] Takcı H., “Centroid sınıflayıcılar yardımıyla meme kanseri teşhisi”, *Gazi Üniversitesi Mühendislik Mimarlık*

Fakültesi Dergisi 31(2), (2016), pp: 323 - 330.

[11] Akyol K., "Meme Kanseri Tanısı İçin Özniteliklerin Öneminin Değerlendirilmesi Üzerine Bir Çalışma", *Academic Platform Journal of Engineering and Smart Systems*, 6(2), (2018), pp:109-115.

[12] Karaci, A. (2020). Predicting Breast Cancer with Deep Neural Networks. In: Hemanth, D., Kose, U. (eds) Artificial Intelligence and Applied Mathematics in Engineering Problems. ICAIAME 2019. Lecture Notes on Data Engineering and Communications Technologies, vol 43. Springer, Cham. https://doi.org/10.1007/978-3-030-36178-5_88.

[13] Kör, H. "Classification of Breast Cancer by Machine Learning Methods", 4th International Symposium on Innovative Approaches in Engineering and Natural Sciences, 2019, pp:508-511.

[14] Yavuz, E. and Eyüpoğlu C., "Meme Kanseri Teşhisi İçin Yeni Bir Skor Füzyon Yaklaşımı" *Düzce Üniversitesi Bilim ve Teknoloji Dergisi* 7(3), (2019) pp: 1045-1060.

[15] Sevli O., "Göğüslerden gelende farklı makine öğrenme tekniklerinin performans karşılaştırması", *Avrupa Bilim ve Teknoloji Dergisi* 16 (2019) pp: 176-185.

[16] Cengil E. and Çınar A., "Göğüs Verileri Metrikleri Üzerinden Kanser Sınıflandırılması" *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, 11(2), (2020) pp: 513-519.

[17] Akcan F. and Sertbaş A., "Topluluk Öğrenmesi Yöntemleri ile Göğüs Kanseri Teşhisi", *Electronic Turkish Studies*, 16(2), (2021), pp: 511 - 527.

[18] Toraman S. and Turkoglu I., "A new method for classifying colon cancer patients and healthy people from FTIR signals using wavelet transform and machine learning techniques", *Journal of the Faculty of Engineering and Architecture of Gazi University*, 35(2), (2020) pp: 933-942.

[19] Breiman L., "Random forests", *Machine Learning*, 45 (1) (2001), pp: 5-32.

[20] Akkurt A., et al., "Developments in the Turkish banking sector: 1980–1990", Issues in Banking Structure and Competition in a Changing World, Conference Proceedings. Central Bank of the Republic of Turkey, Ankara, Turkey. 1992.

[21] Cortes C., ve Vapnik V., "Support-vector networks", *Machine Learning*, 20(3), (1995), pp:273-297.

[22] Platt J., "Sequential minimal optimization: A fast algorithm for training support vector machines", (1998).

[23] Louppe G., "Understanding random forests", Cornell University Library 10 (2014).