Araştırma Makalesi

ENHANCING DETECTION METHOD OF BREAST CANCER USING COIMBRA DATASET

ADIL HANI ABDULKAREEM[†], **Mustafa Cem KASAPBAŞI**[††]
[†] İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, Türkiye,
[††]İstanbul Ticaret Üniversitesi, Mühendislik Fakültesi, İstanbul, Türkiye,
https://orcid.org/0000-0001-8800-1865, https://orcid.org/0000-0001-6444-6659
adil_alnakeeb@yahoo.com, mckasapbasi@ticaret.edu.tr

ABSTRACT

Breast cancer is one of the most dangerous and second most common types of cancer in the world. Breast cancer-fighting with developed devices and medical therapies has become easier. To obtain the best result in breast cancer treatment, periodic checks should be carried out to follow the early diagnosis. Data Mining techniques are used to predict the success of treatment or diagnosis. In this study, the K-Nearest Neighbor (k-NN), Naïve Bayes classifier algorithms of machine learning were used for early detection of breast cancer. From the UC Irvine Machine Learning Repository (UCI) library Coimbra Breast Cancer data set which consists of age, glucose, body mass index (BMI), resistin, insulin, adiponectin, homeostatic model assessment (HOMA), monocyte chemoattractant protein-1 (MCP1), and leptin attributes were used. K-NN model using Age, Resistin, Glucose, and BMI give the highest results, where 90% of specificity 84% percent of sensitivity, and 87.5% accuracy is achieved. These findings provide promising evidence that models combining resistin, glucose, age, and BMI may be a powerful tool for breast cancer detection.

Keywords: Coimbra dataset, Cancer biomarker, Data mining, Breast cancer

COIMBRA VERİ SETİNİ KULLANARAK MEME KANSERİ TESPİT YÖNTEMİNİN İYİLEŞTİRİLMESİ

ÖZET

Meme kanseri, dünyadaki en tehlikeli ve ikinci en yaygın kanser türlerinden biridir. Gelişmiş cihazlarla ve tıbbi tedavilerle meme kanseri ile mücadele daha kolay hale geldi. Meme kanseri tedavisinde en iyi sonucu elde etmek ve erken tanı için periyodik kontroller yapılmalıdır. Makine öğrenme teknikleri, tedavinin başarısını tahmin etmek veya teşhis etmek için kullanılır. Bu çalışmada meme kanserinin erken tespiti için K-En Yakın Komşu (k-NN), Naïve Bayes sınıflandırıcısı makine öğrenmesi algoritmaları kullanılmıştır. Kullanılan veri seti, yaş, glikoz, BMI, resistin, insülin, adiponektin, HOMA, MCP1 ve leptin özelliklerinden oluşan UCI kütüphanesinden alınan Coimbra meme kanser veri setidir. Yaş, Resistin, Glikoz ve BMI kullanan K-En Yakın Komşu modeli en yüksek sonuçları vermektedir. burada özgüllüğün 90%'ı hassasiyetin 84%'ü ve 87.5%'i doğruluk elde edilir. Bu bulgular resistin, glikoz, yaş ve BMI'yı birleştiren modellerin meme kanseri tespiti için güçlü bir araç olabileceğine dair umut verici kanıtlar sunmaktadır.

Anahtar Kelimeler: Coimbra veriset, Bio belirteci, Veri madenciliği, Meme Kanseri

---

Geliş/Received : **07.05.2020**
Gözden Geçirme/Revised : **07.05.2020**
Kabul/Accepted : **31.05.2020**

1. INTRODUCTION

Breast cancer indicates to the abnormal transformation in the cell of the breast tissues. This procedure may form a mass or lump of new tissue which is called tumor. There is two kind of tumor and cancerous (benign) and (malignant). These two types of cancer classification depend on even or not this tumor can spread by metastasis or invasion (Imaginis) (Soliman & AboElHamd, 2014).
According to the American Cancer Society (ACS) study in the United States, a study by 268,600 women who were diagnosed with invasive breast cancer only in 2019, this disease is the male breast cancer that occurs in less than 1% of all breast cancer cases. in America, just in 2019, It is estimated that 41,760 women and 500 men will die because of breast cancer (Society, 2020). Early detection of cancer is one of the important factors for surviving from the illness. Therefore any method or any test that helps physician to diagnose better are vital for human life (Society, 2020).

Finding breast cancer in its early stage is very important thing to prevent death. Health care systems and devices can make use of data mining algorithms which can help to diagnose the diseases in its early stages. Besides clinical and laboratory tests, machine learning algorithms have been widely used to assist doctors in diagnosis (Patrício, et al., 2018). The advantage of machine learning algorithms in medical fields is increasing gradually. Big amount of cancer data was obtained with the coming of new technologies in the field of medicine these data are available now for medical research community.

In this study Breast Cancer Coimbra data set is used (Patrício, et al., 2018), which basically uses ordinary blood test results to be used as a biomarker of breast cancer.

This dataset is chosen since it requires simple blood test markers to predict (UCI). Many other studies also uses the same data set and tries to get better results which are given in the details in literature review section. Interestingly some of our results are better than the original study that is presented in the discussion and evaluation section.

In this study, we used data classification algorithms K-NN, and Naïve Bayes in the search to better predict the presence of the illness. In the human body, there are sundry kinds of tissues formed by a plurality of cells. A human body can develop cancer due to a tumor caused by vertiginous growth and inharmonious of these cells (Brabletz, Jung, Spaderna, Hlubek, & Kirchner, 2005). The results of the treatment linked greatly to the previous recognition of the disease, knowing that can produce severe consequences consequence the delay in the diagnosis (Veer, et al., 2002).

1.1 A Brief Literature Review

Many studies has been carried out using the same data set with different methodologies, a brief of which has been given in this section.

A study by E. Yavuz and C. Eyupoglu. In that study, researchers employed a generalized regression neural network (GRNN). GRNN is considered one of the two variants of the Radial basis function (RBF) network, to classify patterns. Where the data was splitting of 75% training and 25% test. In that study, the algorithm has been implemented in MATLAB software. The researchers calculated the rate of accuracy using this algorithm. The classification accuracy rate of 0.9773 was attained using the GRNN model (Yavuz & Eyupoglu , 2020).
In the research by S.B.Akben. Proposed a decision tree-based method for breast cancer diagnosis. In that study, the researcher did the optimum number of nodes and division algorithm was determined first. Where His study presented a range of boundaries or values indicating cancer owing to the nature of the decision tree, and obtained a relatively accuracy rate of 90.5% (S.B.Akben, 2019).

In the work introduced Y. Li and Z. Chen. Compared and examined the decision tree, logistic regression, performance of artificial neural networks (ANN), random forests (RF), and SVM classifiers for breast cancer prediction. The results of their study showed that the RF algorithm is best than other classification algorithms with an accuracy rate of 74% (Li & Chen, 2018).

A study by M. Patrício et al. Where predictive models were built with classification algorithms SVM, random forests (RF), and logistic regression (LR) implemented in R (R 3.0.2). In that study, researchers the Gini coefficient for the multivariate analysis has used for determining the importance of breast cancer predictors of each of the variables, and power analysis were conducted following the approach described with a few adaptations. Then the researchers calculated the sensitivity, specificity, and AUC. It is observed that SVM models using Resistin, BMI, Glucose, and age was showed better rates. They obtained a specificity ranging from 85% to 90%, sensitivity ranging from 82% to 88% and the 95% confidence interval for the AUC was [0.87, 0.91] (Patrício, et al., 2018)

A study by S. Poorani at el. Deep Neural Network (DNN) classifiers with different hidden layers with different nodes were used to predict breast cancer. Then machine learning algorithms like Decision tree and Support Vector Machine were also trained with the same data. In that study, researchers calculated the rate of accuracy. It is observed that Support Vector Machine, Decision tree, and DNN showed accuracy rates 65.96%, 40.83%, and 75.94% respectively (Poorani & Balasubramanie, 2019).

V. Jonathan Silva Araújo et al. Presented a comparative study used a hybrid artificial intelligence model (fuzzy neural network FNN) and classification algorithms on breast cancer datasets by using WEKA data mining tool. Among the models chosen was the Multilayer Perceptron, decision tree model called C 4.5, Naive Bayes, Zero R, and Random Tree. 70% of data is used for training and 30% is used for testing. The models with the fuzzy neural network obtained were better than the other algorithms with an accuracy of 81%, AUC of 0.8052, sensitivity of 82% and specificity of 81% (Araújo, Guimarães, Souza, Rezende, & Araújo, 2019).

A study by M. F. Aslan, Y. Celik, K. Sabanci, A. Durdu. The algorithms used as Support Vector Machine(SVM), Artificial Neural Network(ANN), K-Nearest Neighbor(KNN), and standard Extreme Learning Machine(ELM). In that study, researchers used the Feature Scaling method for normalization. Where 80% percent of the data were used in the test stage and 20% in the training stage. and have used MATLAB GUI environment for classification in that study. researchers calculated the rate of accuracy rates. The classification accuracy of the ELM obtained by their study was better than the other algorithms. It is observed that the ELM showed accuracy rates of 80%, ANN showed accuracy rates of 79.4%, KNN showed accuracy rates of 77.5% and SVM showed accuracy rates of 73.5% (Aslan, Celik, Sabanci, & Durdu, 2018).

A study by E. Purwaningsih. Compares the performance of Neural Network based on Particle Swarm Optimization (PSO) and Support Vector Machine (SVM) on breast cancer data set. In that study, researchers calculated the rate of accuracy and AUC using those algorithms. In that study, it is observed that the PSO-based Neural Network model has a higher accuracy value of 84.55% from Particle Swarm Optimization based SVM (Purwaningsih, 2019).

T. Kartbayev et al. applied a comparative study on data mining algorithms used in the biomedical field by using WEKA data mining tools. In their research, they implemented different data mining methods which included decision trees (J48), Multilayer Perceptron (MLP), K-NN, and SVM algorithm. In that study, the J48 algorithm with 76.92%, has registered the highest accuracy rate (Gültepe & Kartbayev, 2019).

2. MATERIAL

In this section Comibra data set is going to be described (UCI). In order to better understand the effect of the attributes on the classification decision sub sets of the attributes are gathered. Results of theses subset are compared with each other and discussed in the next section. This section is dedicated for brief description of the assessment methods used in the study.

Despite there are several diverse techniques for the disclosure of breast cancer but the detection trouble of this disease still exists. The data set is obtained from (UCI) consist of 9 quantitative attributes and a label attribute indicating if clinical result is positive for existing of cancer are given in Table 1. As it is stated in the original study the attributes are anthropometric data, which can be gathered in routine blood analysis. Prediction models based on these attributes, if accurate, can potentially be used as a biomarker of breast cancer. These features were measured from 52 healthy people and 64 patients with breast cancer (UCI).

Table 1. Attributes and values

Attribute #	Attribute Name	Value
A1	Age	Numerical Value
A2	BMI (Body Mass Index)	(kg/m^2),
A3	Glucose	(mg/dL)
A4	Insulin	(µU/mL)
A5	HOMA (Homeostatic Model Assessment)	Numerical Value
A6	Leptin	(ng/mL)
A7	Adiponectin	(µg/mL)
A8	Resistin	(ng/mL)
A9	MCP-1	(pg/dL)
Class	Class	{1-Healthy, 2- Patient}

3. METHOD

This section is dedicated for brief description of the assessment methods used in the study.

3.1. Applied Data Mining Methods

As indicated earlier section, in order to comprehend the effects of the attributes subset of these attributes are gathered and used with in the experiments. R Studio is used to perform test and k-fold cross validation is used in validation of the results. A preliminary step for the multivariate analysis consisted of preprocessing the Breast Cancer dataset by deleting records that are duplicates and missing values. The dataset is split randomly into two separate subsets. %20 percent of the dataset is reserved for testing and obtaining the accuracy, sensitivity and selectivity of the methods against the dataset.

3.2. K-Nearest Neighbors (k-NN)

k-NN algorithm is one of the classification algorithms it did not get popular until the 1960s although it was first discovered in the 1950s (Suryawanshi & Sharma, 2016). It is especially used in optical character recognition, pattern recognition, facial recognition, etc. (Han, Kamber, & Pei, 2006). The decision is made in order to classify the data by looking at the dimension from k numbered neighbor (Suguna & Thanushkodi, 2010). In the kNN algorithm, there are two important issues: The first issue is the choice right k which has significant importance on the performance of the algorithm and the second issue is to calculate the dimension between the test instance and its neighbors (Zhang, 2016).

3.2.1. Dimension Measures

The KNN classifier is one of the simplest and most common classifiers where the core of this classifier mainly depends on measuring the distance between the training paradigms and the tested paradigms. Distance measurement is the key to finding the closest neighbor to a point. There are several types of dimension measures such as are the Euclidean, the Minkowski, and the Manhattan dimension, Gini etc. The Euclidean dimension is more generally used than the Manhattan and the Minkowski dimension (Annasaheb & Verma, 2016).
$X_i = (X_1, X_2, X_3, X_4 \dots \dots X_n)$ and $Y_i = (Y_1, Y_2, Y_3, Y_4 \dots \dots Y_n)$ represent feature vectors.
n is the dimension of feature space.

The Euclidean dimension can be calculated as given Eq. (1):

$$\mathrm{D_{Euclidean}}(\mathrm{X},\mathrm{Y}) = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{X_i} - \mathrm{Y_i})^2} \tag{1}$$

The Manhattan dimension can be calculated: as given Eq. (2)

$$\mathrm{D_{Manhattan}}(\mathrm{X},\mathrm{Y}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{X_i} - \mathrm{Y_i}| \tag{2}$$

The Minkowski dimension can be calculated as given Eq. (3):

$$\mathrm{D_{Minkowski}}(\mathrm{X},\mathrm{Y}) = \left(\sum_{\mathrm{i}=1}^{\mathrm{n}} |(\mathrm{X_i} - \mathrm{Y_i})|^{\mathrm{p}}\right)^{1/\mathrm{p}} \tag{3}$$

For a fixed $p \in (0, \infty)$.

3.3. Naïve Bayes

3.3.1. Bayes Theorem

It is one of the methods of algorithms for data classification, which is characterized by its simplicity and effectiveness. Bayes theorem has an important role in probability calculations. It is possible to classify based on Bayes theorem, which is discovered by a mathematician Thomas Bayes, in the 18th century (Özkan, 2016) (URAL, 1978) and can be calculated as given Eq. (4).

$$P(H|X) = \frac{p(X|H)P(H)}{P(X)} \tag{4}$$

$P(H|X)$ is the posterior probability and means the probability of the hypothesis occurring by knowing that the evidence has already occurred.

$p(X|H)$ is the probability that the evidence will occur by knowing that the hypothesis has actually occurred.

$P(H)$ is the probability that the hypothesis will occur.

$P(X)$ is the probability that the evidence will occur.

3.3.2. Gaussian Naive Bayes classifier

A Gaussian distribution is also known as Normal distribution. It is better if we utilized it in conditions when all features are continuous (Sammut & Webb, 2017) (Han, Kamber, & Pei, 2006). Assuming that the numerical data distribution is normal and if the attribute values are numerical, the following standard probability density function is used (Özkan, 2016) and can be calculated as given in Eq. (5).

$$P(x_k|C_i) = f\left(x_k, \mu_{c_i}, \sigma_{c_i}\right) = \frac{1}{\sqrt{2\pi} \cdot \sigma_{c_i}} e^{-\frac{\left(x_k - \mu_{c_i}\right)^2}{2\sigma_{c_i}^2}} \tag{5}$$

Here, μ_{c_i} is the average and σ_{c_i} is the standard deviation.

4. Result and Discussions

In this work, the dataset was taken from the UCI library (UC Irvine Machine Learning Repository). Breast Cancer Coimbra data set (UCI) were used where there are 116 samples in a total of 64 patients with breast cancer and 52 healthy controls.

KNN and NB classification algorithms are applied to the Breast Cancer Coimbra data set. Result of the applied kNN and Naïve Bayes are given in the last four rows of Table 2 in comparison with earlier studies using the same data set. The main aim is to build a classification model that allows us to compare the performance of these two methods with other studies are carried out according to. Sensitivity, specificity, AUC and accuracy are used as criteria for assessing performances based on potential biomarkers from the nine predictors. Experiments and code implementation parts are implemented by using R Studio. Training and test data were generated randomly from the data where 80% percent were used in the test phase and 20% percent of the whole data were used in the training phase.

What distinguishes our study that instead of using the entire data sets like most of the previous studies. we used the first four biomarkers using resistin, glucose, age, and BMI as biomarkers to predict the presence of breast cancer. As a result of previous studies (Patrício, et al., 2018) and after conducting several experiments on our study. We obtained better and finer results when using these biomarkers.

Table 2. Attributes and values in the table

Methods	Sensitivity	Specificity	AUC	Accuracy
SVM (Patrício, et al., 2018)	0.81	0.84	0.85	-----
RF (Patrício, et al., 2018)	0.85	0.77	0.83	-----
LR (Patrício, et al., 2018)	0.76	0.86	0.81	-----
ANN (Aslan, Celik, Sabanci, & Durdu, 2018)	-----	-----	-----	0.794
ELM (Aslan, Celik, Sabanci, & Durdu, 2018)	-----	-----	-----	0.80
KNN (Aslan, Celik, Sabanci, & Durdu, 2018)	-----	-----	-----	0.775
SVM (Aslan, Celik, Sabanci, & Durdu, 2018)	-----	-----	-----	0.735
ANN AndNet (Araújo, Guimarães, Souza, Rezende, & Araújo, 2019)	0.765	0.245	0.51	0.53
ANN OrNet (Araújo, Guimarães, Souza, Rezende, & Araújo, 2019)	0.76	0.459	0.63	0.628
ANN UniNet (Araújo, Guimarães, Souza, Rezende, & Araújo, 2019)	0.783	0.412	0.62	0.62
DNN (Poorani & Balasubramanie, 2019)	-----	-----	-----	0.759
SVM-PSO (Purwaningsih, 2019)	-----	-----	0.819	0.81
NN-PSO (Purwaningsih, 2019)	-----	-----	0.885	0.845
J48 (Gültepe & Kartbayev, 2019)	-----	-----	-----	0.769
MLP (Gültepe & Kartbayev, 2019)	-----	-----	-----	0.692
KNN (Gültepe & Kartbayev, 2019)	-----	-----	-----	0.692
SVM (Gültepe & Kartbayev, 2019)	-----	-----	-----	0.664
kNN A1-A9 (all) k=7	**0.7692**	**0.9091**	**0.839**	**0.8333**
kNN A1-A4 (selected) k=7	**0.8462**	**0.9091**	**0.8776**	**0.875**
Naive Bayes A1-A9 (all)	**0.7**	**0.6667**	**0.68**	**0.6818**
Naive Bayes A1-A4 (selected)	**0.75**	**0.83**	**0.794**	**0.79**

While using the KNN algorithm, it is important to decide the right k value. To see the performance of the KNN classification model on the Breast Cancer Coimbra data set, it is built a confusion matrix as shown in Table 3 were 92 records used to train the algorithms and 24 records used to test the classification algorithms.

Table 3. KNN algorithm confusion matrix

	Positive 1	Negative 2
Positive 1	11	1
Negative 2	2	10

k=7 gives higher accuracy than the other k values with 87.5%, Sensitivity 84.62%, Specificity 91%, and AUC 87.76%, according to table 2.

Naive Bayes algorithm calculates the probability using the given data so it is known as a fast and efficient classification method. In this study, the NB algorithm is calculated based on Gaussian distribution. According to the confusion matrix from Table 5, the accuracy of a Naive Bayes algorithm as 79%. Sensitivity 75%, Specificity 83%, and AUC 79%, according to Table 2.

Table 5. Naive Bayes algorithm confusion matrix

	Positive 1	Negative 2
Positive 1	10	2
Negative 2	3	9

Data set is partitioned in two according to attributes and the same experiments are carried out for all attributes included and the first four attributes included resistin, glucose, age, and BMI.

Result of the applied kNN and Naïve Bayes are given in last four rows of the Table 2 in comparison with earlier studies that of same data set. Comparisons with other studies are carried out according to Sensitivity, Specificity, AUC and Accuracy. However in some of the compared studies not all the metrics were given.

the KNN algorithm has the best accuracy rate comparing to the NB algorithm where it reached an 87.5% accuracy rate.

These findings provide promising evidence that models combining resistin, glucose, age, and BMI may be a powerful tool for breast cancer detection.

5. CONCLUSION

In this study a well-known breast cancer data set is used in conjunction with two machine learning algorithms namely kNN and Naïve Bayes. Based on the first four biomarkers Age, Glucose, Resistin, and BMI. The proposed BC logistic model has shown that it can be used to detect breast cancer. With kNN, 90% of specificity 84% percent of Sensitivity, and 87.5% accuracy is achieved. When compared with the other studies this study indicate that Its results are good than compared regarding the same data set. Besides having good results kNN, and Naïve Bayes are relatively easier methodologies to develop. If these results are used in assessing the existence of benign or malignant cases such a method will aid physicians to diagnose better for breast cancer early detection.

REFERENCES

Annasaheb, A. B., & Verma, V. K. (2016). Data Mining Classification Techniques: A Recent. International Journal of Emerging Technologies in Engineering Research, 51-54.

Araújo, V. J., Guimarães, A. J., Souza, P. V., Rezende, T. S., & Araújo, V. S. (2019). Using Resistin, Glucose, Age and BMI and Pruning Fuzzy Neural Network for the Construction of Expert Systems in the Prediction of Breast Cancer. Machine Learning and Knowledge Extraction, 466-482.

Aslan, M. F., Celik, Y., Sabanci, K., & Durdu, A. (2018). Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis Data. International Journal of Intelligent Systems and Applications in Engineering, 289-293.

Brabletz, T., Jung, A., Spaderna, S., Hlubek, F., & Kirchner, T. (2005). Migrating cancer stem cells — an integrated concept of malignant tumour progression. Nat Rev Cancer, 744-749.

Gültepe, Y., & Kartbayev, T. (2019). A Study of Data Mining Methods for Breast Cancer Prediction. Proceedings Book, 303-305.

Han, J., Kamber, M., & Pei, J. (2006). Data Mining: Concepts And Techniques. San Francisco: University of Illinois.

Imaginis. (n.d.). Retrieved March 2019, from http://www.imaginis.com/general-information-on-breast-cancer/what-is-breast-cancer-2

Li, Y., & Chen, Z. (2018). Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction . Applied and Computational Mathematics, 212-216.

Özkan, Y. (2016). Veri Madenciliği Yöntemleri. Istanbul: Papatya Yayıncılık Eğitim.

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., et al. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer, 1-8.

Poorani, S., & Balasubramanie, P. (2019). Deep Neural Network Classifier in Breast Cancer Prediction. International Journal of Engineering and Advanced Technology, 2106-2109.

Purwaningsih, E. (2019). Application of the Support Vector Machine and Neural Network Model Based on Particle Swarm Optimization for Breast Cancer Prediction. SinkrOn, 66-73.

S.B.Akben. (2019). Determination of the Blood, Hormone and Obesity Value Ranges that Indicate the Breast Cancer, Using Data Mining Based Expert System. Innovation and Research in BioMedical engineering IRBM, 355–360.

Sammut, C., & Webb, G. I. (2017). Encyclopedia of Machine Learning and Data Mining. Springer Publishing Company, Incorporated, 2nd edition.

Society, A. C. (2020). Breast Cancer Facts & Figures 2019-2020. A Cancer Journal for Clinicians.

Soliman, O. S., & AboElHamd, E. (2014). Classification of Breast Cancer using Differential Evolution and LeastSquares Support Vector Machine. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 3(2), 155–161.

Suguna, N., & Thanushkodi, K. (2010). An Improved k-Nearest Neighbor Classification Using Genetic Algorithm. International Journal of Computer Science Issues, 18-21.

Suryawanshi, A., & Sharma, A. (2016). A Novel Method for Detecting Spam Email using KNN Classification with Spearman Correlation as Distance Measure. International Journal of Computer Applications , 28-35.

UC Irvine Machine Learning Repository. (n.d.). Retrieved from http://archive.ics.uci.edu/ml/index.php

UCI. (n.d.). Retrieved from Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php

URAL, K. (1978). İstatistik ve Karar Alma. Istanbul: Istanbul University.

Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature, 530–536.

Yavuz, E., & Eyupoglu , C. (2020). An effective approach for breast cancer diagnosis based on routine blood analysis features. Med Biol Eng Comput.

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of Translational Medicine, 218-218.