İLERİ MÜHENDİSLİK ÇALIŞMALARI VE TEKNOLOJİLERİ DERGİSİ

Meme Kanseri Tespitinde Sınıflandırma ve Sinir Ağları Yöntemlerinin Karşılaştırılması

Elif ÖZTAD*[1]

*İstanbul Aydın Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Ana Bilim Dalı, İstanbul, 34295, Türkiye

Araştırma Makalesi, Geliş Tarihi: 24.07.2020, Kabul Tarihi: 17.08.2020

Özet

Meme kanseri, çoğalan ve çoğu zaman tümör adı verilen bir kitle oluşturan bazı hücrelerin bozulmasından kaynaklanır. Tümörler iyi huylu (kanserli olmayan) veya kötü huylu (kanserli) olabilmektedir. MRG (manyetik rezonans görüntüleme), mamogram, ultrason ve biyopsi gibi testler yaygın olarak yapılan meme kanserini teşhis etmek için kullanılır. Verilerde iki farklı etiket olduğundan, tahmin iki kategoriye ayrılır (Kötü huylu veya iyi huylu). Makine öğreniminde bu bir sınıflandırma problemi olarak tanımlanır. Bu çalışma, meme kanserinin iyi huylu veya kötü huylu olup olmadığını sınıflandırmayı ve belirli bir süre sonra kötü huylu vakaların nüksünü ve nüksünü öngörmeyi amaçlamaktadır. Kullanılan metodoloji sınıflandırma modelini ve sinir ağları metodunu içerir. Python modülleri, verileri iyi bir şekilde kavramak ve verilerin farklı şekillerde nasıl ele alınacağını düşünmek için verileri tanımak amacıyla harici veri kümelerini içe aktarmak için kullanılmaktadır. Bu amaçla veri setinden makine öğreniminin temel kavramları uygulanır ve sonuçlar veri setine göre değerlendirilir. Bu nedenle, meme kanseri tahmininde yüksek bir doğruluk elde etmek için sinir ağı yöntemleri ve sınıflandırma yöntemleri birbirleriyle karşılaştırılmaktadır.

Anahtar Kelimeler: Meme kanseri, Sinir ağları, Derin öğrenme, Sınıflandırma.

Comparision Between Classification and Neural Network Methods In Breast Cancer Detection

Abstract

Breast cancer is caused by the breakdown of some cells that multiply and often form a mass called a tumor. Tumors can be benign (not cancerous) or malignant (cancerous). Tests such as MRI (magnetic resonance imaging), mammogram, ultrasound, and biopsy are used to diagnose common breast cancer. Since there are two different labels in the data, the estimate is divided into two categories (benign or malignant). In machine learning, this is defined as a classification problem. This study aims to classify whether breast cancer is benign or malignant and predict the relapse and relapse of malignant cases after a certain period of time. The methodology used includes the classification model and neural network methods. Python modules are used to import external datasets in order to grasp the data well and to think about how to handle the data in different ways. For this purpose, basic concepts of machine learning are applied from the dataset and the results are evaluated according to the dataset. Therefore, neural network methods and classification methods are compared with each other to achieve a high accuracy in breast cancer estimation.

Keywords: Breast cancer, Neural networks, Deep learning, Classification.

[1]Sorumlu yazar elifoztad@stu.aydin.edu.tr

1. GİRİŞ

Meme kanserinin kadınlarda en sık teşhis edilen kanser türü olduğu bilinmektedir. Sadece kadınları değil, çok az bir oranda da olsa erkekleri de etkileyebilmektedir. Bu konuda uzmanlardan biri olan Profesör Aghzadi Rajaa'ya göre, her 10 kadından biri hayatı boyunca meme kanseri riski altındadır. Bu da meme kanserini büyük bir halk sağlığı sorunu haline getirmeye yetmektedir. Meme kanseri için birçok etkili tedavi olmasına rağmen, tedavi sonrasında önemli sayıda hastada hastalık yeniden nüksedebilmektedir ve bu durum yeni tedavi stratejilerinin araştırılmasını sağlamıştır (Ouardirhi, 2020).

Hızla gelişen teknoloji, meme kanseri için hayati önem taşıyan erken teşhisin yapılmasına katkı sağlamış, sınıflandırma algoritmaları ve derin öğrenme ile yüksek doğruluk elde edebilen tahminlerin ortaya çıktığı projeler geliştirilmiştir. Veri madenciliği yöntemlerinden yararlanarak yapılan tespit, uygulanan tedavilerin agresifliğini önemli ölçüde azaltabilmektedir.

Bu çalışmada, meme kanserinde tümörün iyi huylu veya kötü huylu olup olmadığını sınıflandırmak ve belirli bir süre sonra kötü huylu vakaların nüksünü öngörmek amaçlanmıştır.

2. KONULAR

2.1. Meme Kanseri Tespitinde Sınıflandırma Algoritmalarının Kullanılması

Sınıflandırma, kategori üyeliği makine öğrenimi ve istatistiklerinde bilinen gözlemleri (veya örnekleri) içeren bir veri seti temelinde, yeni bir gözlemin hangi kategorilerden hangisine ait olduğunu belirleme sorunudur (Rana, 2015). Bu çalışmada sınıflandırma algoritmaları kullanılmıştır. Bu algoritmalar aşağıda belirtilmiştir.

K-En Yakın Komşu: Bu algoritma, basit olması ve düşük hata oranına sahip olması ile bilinmektedir. Öğrenme için mevcut olan en basit sınıflandırma algoritmalarından biridir. Amaç, özellik alanındaki test verilerinin en yakın eşleşmesini aramaktır (Rana, 2015).

Karar Ağaçları: Bu tür modeller insanın akıl yürütmesine çok benzediği ve anlaşılması kolay olduğu için sınıflandırma modelleri oluşturmak için yaygın olarak kullanılmaktadır (Kotsiantis, 2013).

DVM – Destek Vektör Makineleri (DVM): Nesnelere etiket atamayı örnek olarak öğrenen bir bilgisayar algoritmasıdır. Örneğin, yüzlerce veya binlerce hileli ve hileli olmayan kredi kartı aktivite raporlarını inceleyerek hileli kredi kartı etkinliğini tanımayı öğrenebilir (Noble, 2006).

Her biri iki kategoriden birine ait olarak işaretlenmiş bir dizi eğitim örneği göz önüne alındığında, bir DVM eğitim algoritması bir kategoriye veya diğerine yeni örnekler atayarak bir olasılık dışı ikili doğrusal sınıflandırıcı haline getirir (Garg, 2018).

Naif Bayes (NB): Naif Bayes Algoritması, her özellik çifti arasında bağımsızlık varsayımı ile Bayes Teoremi'ne dayanmaktadır. Naif Bayes sınıflandırıcılar, belge sınıflandırması ve spam filtreleme gibi birçok koşulda iyi çalışır. Naif Bayes sınıflandırıcıları, daha sofistike yöntemlere kıyasla son derece hızlıdır.

Lojistik Regresyon (LR): Lojistik Regresyon, sınıflandırma için bir makine öğrenme algoritmasıdır. Bu algoritmada, tek bir denemenin olası sonuçlarını tanımlayan olasılıklar bir lojistik fonksiyon kullanılarak modellenmiştir ve bu amaç için tasarlanmıştır. Ayrıca birçok bağımsız değişkenin tek bir sonuç değişkeni üzerindeki etkisini anlamak için en kullanışlı olan algoritmadır (Rouse, 2018). Lojistik regresyon, 1 ve 0 değerlerini alan bir ikili yanıt değişkenini modellemek için bir yöntem sağlar (Bewick, Cheek ve Ball, 2005).

Evrişimsel Sinir Ağları (ESA): İngilizce'de Convolutional Neural Network (CNN) olarak geçen bu algoritma, derin öğrenmede kullanılmaktadır. Diğer algoritmalardan farklı olarak, insan beyninde istemsiz olarak gerçekleşen, objeleri en ince ayrıntısına kadar inceleme ve bağıntı kurma özelliğini taklit eder. Örneğin insanlara Medusa Heykeli gösterildiğinde bu heykelin Medusa Heykeli olduğunu başındaki yılanlardan anlarlar, yılan ve baş objelerini birleştirip Medusa olduğuna karar verirler. Derin öğrenmede insanların bu özelliği taklit edilerek, örneğin; basketbol topu, basketbol ayakkabıları giymiş biri ve potanın olduğu fotoğraftan, o kişinin basketbol oynadığı çıkarımında bulunulabilir. Objelerin birbirleri ile ilişki ve bağlantıları ortaya konur. Bu algoritmanın 5 katmanı mevcuttur. Katmanları kısaca aşağıdaki gibidir.

Evrişimli Katman: Özellikleri saptamak için kullanılmaktadır (Ergin, 2018).

Doğrusal Olmayan Katman: Sisteme doğrusal olmayanlığı (non-linearity) tanıtmaktadır (Ergin, 2018).

data = pd.read_csv('data/data.csv', index_col=False,)

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	r
0	842302	M	17.99	10.38	122.8	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.9	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.0	1203.0	0.10960	0.15990	0.1974	0.12790	...	

Şekil 1. Veri setinden alınan bir kesit

Havuzlama (Altörnekleme) Katmanı: Ağırlık sayısını azaltır ve uygunluğu kontrol etmektedir (Ergin, 2018).

Düzleştirme Katmanı: Klasik Sinir Ağı için verileri hazırlamaktadır (Ergin, 2018).

Tamamen Bağlı Katman: Sınıflandırmada kullanılmaktadır (Ergin, 2018).

2.2. Meme Kanseri Tespitinde Sinir Ağı Yöntemlerini Kullanma

Bilgi teknolojisinde (BT), bir sinir ağı, insan beynindeki nöronların çalışmasından sonra şekillendirilen bir donanım ve yazılım sistemidir. Sinir ağları uyarlanabilir olmaları açısından dikkat çekicidir, yani başlangıç eğitiminden öğrendiklerinde kendilerini değiştirirler ve daha sonraki çalışmalar dünya hakkında daha fazla bilgi sağlar (Chi, Street ve Wolberg, 2007). En temel öğrenme modeli girdi akışlarını ağırlıklandırmaya odaklanmıştır, bu da her bir düğümün öncüllerinden gelen girdinin önemini nasıl ağırlaştırdığıdır (Chi, Street ve Wolberg, 2007). Doğru cevapların alınmasına katkıda bulunan girdiler daha ağırdır (Chi, Street ve Wolberg, 2007).

3. ANALİZ METODU

Bu projedeki temel amaç, belirli bir süre sonra kötü huylu vakaların nüksünü tahmin etmektir. Dolayısıyla, tahakkümün burada önemli bir başlangıç noktası vardır. Makineyi eğitmek için meme kanseri tümörünün veri kümesine ihtiyaç vardır. Bu nedenle, meme kanseri veri kümesi, California Üniversitesi tarafından tutulan makine öğrenimi deposundan alınmıştır.

3.1. Veri Kümesinde Bulunan Bilgi Türlerini Tanımlama

Bu projede, verileri iyi bir şekilde kavramak ve verilerin farklı şekillerde nasıl ele alınacağını düşünmek için verileri tanımak amacıyla harici veri kümelerini içe aktarmak için Python modülleri kullanılmaktadır. Şekil 1'de veri setinden alınan bir kesit gösterilmiştir.

3.2. Keşifsel Veri Analizi

Veri keşfi ve görselleştirme teknikleri Python kütüphanelerinde (Pandas, matplotlib, deniz dibi) kullanılmaktadır. Şekil 2'de kullanılan kütühanelerin kod görüntüsünden bir kesit gösterilmiştir.

```
import numpy as np
import pandas as pd
from scipy.stats import norm
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.model_selection import cross_val_score
from sklearn.pipeline import make_pipeline
from sklearn.metrics import confusion_matrix
from sklearn import metrics, preprocessing
from sklearn.metrics import classification_report
from sklearn.cross_validation import cross_val_score, KFold
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.pipeline import Pipeline
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression

from sklearn.metrics import accuracy_score
```

Şekil 2. Kullanılan kütüphanelerin kod görüntüsünden bir kesit

SVM algoritma sınıflandırıcısı %70 eğitim datasında kullanıldı.

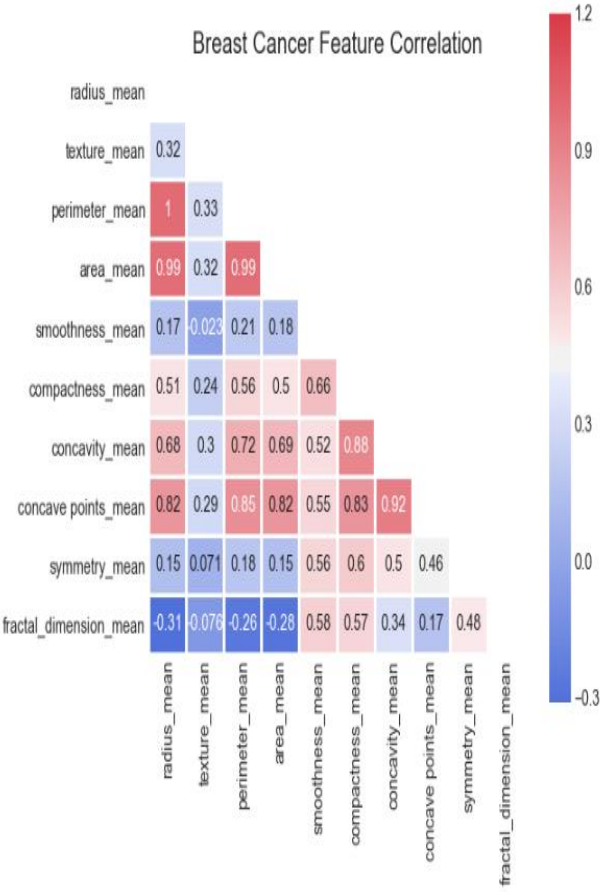
```
clf = SVC(probability=True)
clf.fit(X_train, y_train)
```

30 özellikli bir NumPy dizisinde X, "M" ve "B" dizelerine "1" ve "0" tamsayıları eklendi. Tümör kötü huylu ise, kötü huylu olduğu anlamına gelen "1" yazılacaktır. Aksi takdirde, "0" olarak yazılacak, yani kötü huylu değil, iyi huylu anlamında.

```
array = data.values
X = array[:,1:31]
y = array[:,0]
#transform the class labels from string "M" and "B" to integers
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
y = le.fit_transform(y)
#Call the transform method of LabelEncorder on two dummy variables
le.transform (['M', 'B'])
#Malignant = 1 (indicates prescence of cancer cells)
#Benign = 0 (indicates abscence)
```

Şekil 3. "M" ve "B" etiketlerinin "1" ve "0" a dönüştürülmesinden bir kesit

Burada string değerleri integer değerlere çevirmek önemlidir çünkü sonuç değeri olarak string kabul edilemez. Şekil 3'te bu etiketlerin "1" ve "0"a dönüştürülmesinden bir kesit gösterilmiştir.



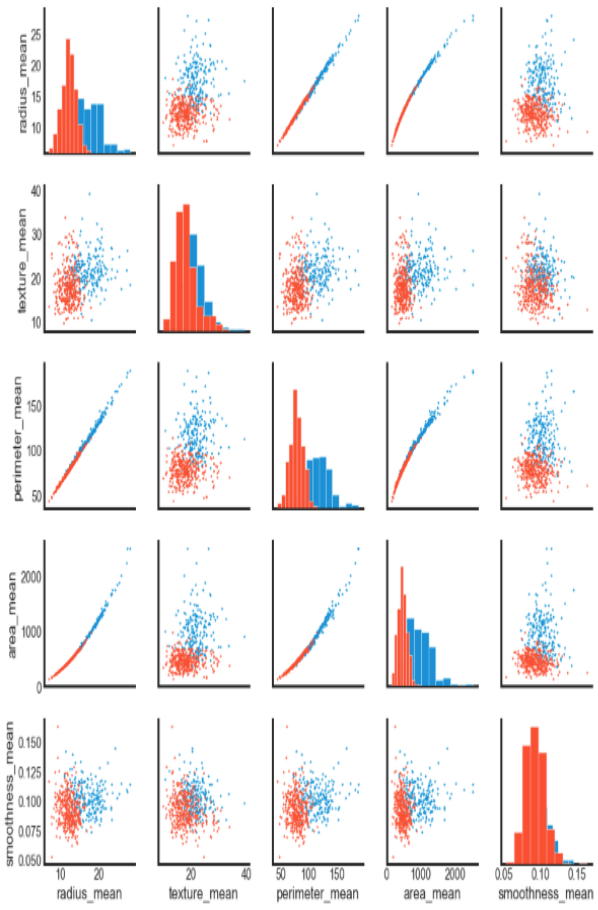
Şekil 4. Korelasyon matrisi

Şekil 4'te, 1-0,75 arasındaki ortalama değer parametrelerinde güçlü pozitif ilişkinin olduğu görülmektedir.

• Doku çekirdeğinin ortalama alanı, yarıçap ve parametrenin ortalama değerleri ile güçlü bir pozitif korelasyona sahiptir.

• İçbükeylik ve alan gibi bazı parametreler orta derecede pozitif ilişkilidir.

• Benzer şekilde, fraktal boyut ile yarıçap, doku, parametre ortalama değerleri arasında güçlü bir negatif korelasyon görülmektedir. Aşağıda şekil 5'te kırmızı renk kötü huylu, mavi renk iyi huylu tümörü temsil etmek üzere histogram gösterilmiştir.



Şekil 5. Histogram

• Kanserin sınıflandırılmasında hücre yarıçapı, çevre, alan, kompaktlık, içbükeylik ve içbükey noktaların ortalama değerleri kullanılabilmektedir. Bu parametrelerin daha büyük değerleri, kötü huylu tümörler ile bir korelasyon gösterme eğilimindedir.

• Doku, pürüzsüzlük, simetri veya fraktüel boyutun ortalama değerleri, bir tanının diğerine göre belirli bir tercihini göstermez.

• Histogramların herhangi birinde, daha fazla temizliği garanti eden fark edilir büyük uç değerler yoktur.

4. ANALİZ

Şekil 6'daki kod görüntüsü sınıflandırma algoritmalarının doğruluk karşılaştırılmasına aittir. Şekil 7'deki kod görüntüsü ise CNN algoritmasına aittir.

```
models = []
models.append(( 'LR' , LogisticRegression()))
models.append(( 'KNN' , KNeighborsClassifier()))
#models.append(( 'DecisionTree' , DecisionTreeClassifier()))
#models.append(( 'NB' , GaussianNB()))
models.append(( 'SVM' , SVC()))
num_folds = 10
num_instances = len(X_train)
seed = 7
scoring = 'accuracy'
num_folds = 10
num_instances = len(X_train)
seed = 7
scoring = 'accuracy'
results = []
names = []
print("COMPARE ACCURACY RESULTS ")
for name, model in models:
 kfold = KFold(n=num_instances, n_folds=num_folds, random_state=seed)
 cv_results = cross_val_score(model, X_train, y_train, cv=kfold, scoring=scoring)
 results.append(cv_results)
 names.append(name)
 msg = "%s: %f" % (name, cv_results.mean())
 print(msg)
```

Şekil 6. Sınıflandırma algoritmaları kod görüntüsü

```
y_scores = net.predict(test_X)

y_scores = [np.argmax(x) for x in y_scores]

accuracy = accuracy_score(y_true, y_scores)

# inference

print("Network accuracy: {}".format(accuracy))



Network accuracy: 0.767857142857
```

Şekil 7. CNN Algoritması kod görüntüsü

5. SONUÇLAR VE TARTIŞMALAR

Meme kanserini önleme ve mücadele açısından halkın bu konuda bilgilendirilmesi büyük önem taşımaktadır. Dolayısıyla, medyanın meme kanserine karşı mücadelede etkili bir rolü olduğu aşikardır. Meme kanserinin erken tespit edilebilmesi iyileşme şansını arttırabileceği için, erken tespit konusunda literatürde veri madenciliği yöntemlerinden yararlanılarak yapılmış pek çok çalışma mevcuttur. Örneğin Oklahoma Eyalet Üniversitesi'nden araştırmacıların konuyla ilgili çalışmasında (Delen, Walker ve Kadam (2004) karar ağacı algoritmasının % 93,6 doğruluk verdiğini ve tahmin konusunda şuana kadar yapılan çalışmalardaki en yüksek doğruluğa sahip olduğu, yapay sinir ağları ile % 91,2 doğruluk ile ikinci en yüksek doğruluk elde ettikleri, lojistik regresyon ile % 89,2 doğruluk elde ettikleri saptanmıştır. Abdelghani ve Güven'in (2004) yaptıkları çalışmada ise C4.5 Algoritması Naif Bayes Algoritması'nı kullanarak 151.886 kayıttan oluşan veriyi Weka Programı'nı kullanarak analiz etmişlerdir. Doğruluk performansları yüksek sonuçlar vermiştir ancak, verilen veriler için C4.5 Algoritması tarafından üretilen modelin diğer iki tekniğe göre çok daha iyi bir performansa sahip olduğunu keşfetmişlerdir.

LR: 0,948450

KNN: 0,932060

DecisionTree: 0,920044

NB: 0,934164

DVM: 0,607918

CNN: 0,767857142857

Burada görüldüğü gibi, Lojistik Regresyon (LR) %94 doğruluk, En Yakın Komşu (KNN) Algoritması %93, Karar Ağacı (DecisionTree) %92, Naif Bayes (NB) %93, Evrişimsel Sinir Ağları (CNN) ise %76 doğruluk sağlarken, Destek Vektör Makineleri (SVM) bu veri seti için diğer algoritmalar kadar yüksek bir doğruluk sağlayamamıştır.

Bu çalışmada test edilen sınıflandırma yöntemleri, iyi huylu veya kötü huylu tümör sayısı seçiminde algoritmaların önemini göstermiştir. Çalışma, LR, KNN (En Yakın Komşu), Karar Ağacı ve NB Algoritmaları'nın eğitim sürecinde başarısını göstermiştir. Doğruluk oldukça yüksektir. Ancak DVM Algoritması'nda çok düşüktür. Son olarak, iyi bir veri seti ve o veri setine uygun doğru algoritmayı seçmek bizlere doğruluğu yüksek sonuçlar verecektir.

KAYNAKLAR

Abdelghani, B. ve Guven, E. (2004). Predicting Breast Cancer Survivability using Data Mining Techniques. Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining.

Bewick, V., Cheek, L. ve Ball, J. (2005) Statistics Review 14: Logistic Regression Crit Care 9, 112. doi:10.1186/cc3045.

Chi, L., Street, W. ve Wolberg, H. (2007). Application of Artificial Neural Network Based Survival Analysis on Two Breast Cancer Datasets- AMIA Annu Symp Proc. 2007,130–134.

Delen, D., Walker, G. ve Kadam, A. (2004). Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods. doi:10.1016/j.artmed.2004.07.002.

Ergin, T. (2018). Convolutional Network (ConvNet ya da CNN) Nedir, Nasıl Çalışır. https://www.mathworks.com/solutions/deep-learning/convolutional-neural-network.html erişim adresi.

Garg, R. (2018). 7 Types classification algorithms, *Analytics India Magazine*. https://www.analyticsindiamag.com/7-types classification-algorithms erişim adresi.

Kotsiantis, S. B. (2013). Decision trees: a recent overview, *Artificial Intelligence Review*. vol. 39, 261–283.

Noble, W. S. (2006). What is a support vector machine. *Nature Biotechnology*, 24(12).

Ouardirhi, A. (2020). Meme kanseri: bilgi, farkındalık ve önleme, *Al Bayane*. http://albayane.press.ma/cancer-du-sein-linformation-la-sensibilisation-et-la-prevention-dabord.html erişim adresi.

Rana, D. (2015). One class SVM vs SVM classification, *International Journal of Science and Research* (IJSR) ISSN (Online): 2319-7064.

Rouse, M. (2018). Artificial Neural Network (ANN). https://searchenterpriseai.techtarget.com/definition/neural-network erişim adresi.