Makine Öğrenmesi ile Göğüs Kanseri Tahmini

Uğurcan YILMAZ
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi,Istanbul, Türkiye
{l1116135}@std.yildiz.edu.tr

***Özetçe* —Bu çalışmada makine öğrenmesi yöntemlerinin geliştirilmesi ile sağlık alanında kullanılması hedeflenmektedir. Kadınları ilgilendiren bir sağlık sorununun çözümü için sürecin başlangıcı her sene gerçekleşen kanser ölümleriydi. Daha önce birçok kişi tarafından kullanılan göğüs kanseri veriseti bu çalışmada da kullanılmıştır. Makine öğrenmesi yöntemlerinin kullanılması amaçlanmaktadır. Amaç kanserli olan verileri tespit etmek ve bu tespitin başarı oranı en iyi olanı bulmaktı. Daha önce yapılmış bazı projeler incelenip hangi modelin daha iyi olabileceği belirlendi. Bunlar denendi ve en başarılı sonuç raporlandı.**

Anahtar Kelimeler—Makine Öğrenmesi, Python, Confusion Matrix ,Google Colab, Numpy, Pandas, Seaborn, Matplotlib, Earlystopping, Cross Validation, GridSearchCV, Logistic Regression, Random Forest.

I. GİRİŞ

Göğüs kanseri, kadınlar arasında en çok görülen hastalıklardan biridir ve riski de oldukça fazladır. Kadınlar için en tehlikeli hastalıklardan biridir. Dünya Sağlık Örgütü'nün verilerine göre her sene yaklaşık iki milyon kadın göğüs kanseri sebebiyle hayatını kaybetmektedir. Bu nedenle de erken teşhis göğüs kanseri hastalığı için çok önemli bir rol oynamaktadır. [1] Gelişen teknolojilerle birlikte bilgisayarlar insanlardan çok daha hızlı ve kolayca kanser teşhisi yapabilmektedir. Daha kaliteli bilgi toplamak ve daha iyi analiz yapabilmek adına son zamanlarda çok daha verimli makine öğrenmesi algoritmaları tasarlanmıştır. Bu algoritmalar bilgisayarların daha zayıf olduğu taraf olan karar verme noktasında daha iyi hale getirilmiştir. Geliştirilen bu algoritmalar sağlık sektörüne epey yardımcı olmaktadır. Göğüs kanserinin de diğer kanser türlerinde olduğu gibi erken teşhis önemli rol oynamaktadır. Tanı koymak için uzman kişilerin test sonuçlarını yorumlaması gerekmektedir. Lakin gelişen makine öğrenmesi uygulamaları da bu teşhis için önemli adımlar atmaktadır. Bu projede 30 adet özellikten ve 569 veriden oluşan sklearn de bulunan breast cancer veri seti kullanılmıştır. Bu çalışmayı yapmadan önce daha evvelden yapılmış projeler incelenmiş olup uygun algoritmaların tespiti yapılmıştır. [2]

İncelediğim projelerden bazıları:

- Naresh Khuriwal ve arkadaşları 2018 yapıtğı çalışmada Makine Öğrenmesi algoritmalarını kullanarak göğüs kanseri teşhisi yapmayı amaçlamaktadır. Wisconsin veri setini kullanarak yapılan bu çalışmada yapay sinir ağları ile birlikte lojistik regresyonun başarısının %96 olduğu bulunmuştur.[3]
- Ebru Aydınak ve arkadaşları 2019 yılında yaptıkları çalışmada göğüs kanseri veri setlerinin makine öğrenmesi algoritmaları ile karşılaştırılması yapılması hedeflenmektedir. Bu çalışmada, Wisconsin Meme Kanseri veri setinin sınıflandırılması için makine öğrenme tekniklerinden ikisi kullanılmış ve bu tekniklerin sınıflandırma performansları doğruluk, kesinlik, callbacks ve ROC Alanı değerleri kullanılarak birbirleriyle karşılaştırılmıştır. En iyi performans, en yüksek doğrulukla Destek Vektör Makinesi tekniği ile bulunmuştur. [4]
- Sharmin Ara ve arkadaşları 2021 yılında yaptığı çalışmada makine öğrenmesi ile iyi huylu, kötü huylu göğüs kanseri sınıflandırmasını yapmayı amaçlamıştır. Bu çalışmada da Wisconsin Meme Kanseri Veri Kümesi kullanılmıştır. Burada, tümörleri iyi huylu ve kötü huylu olarak sınıflandırmak için Destek Vektör Makinesi, Lojistik Regresyon, K-En Yakın Komşular, Karar Ağacı, Naive Bayes ve Random Forest sınıflandırıcıları uygulanmıştır. Analize göre, Rastgele Orman ve Destek Vektör Makinesi, %96,5 doğrulukla diğer sınıflandırıcılardan daha iyi performans gösteriyor. [5]
- Anuj mangal ve arkadaşları 2021 yılında yaptıkları bu öalışmada makine öğrenmesi tekniklerini kullanarak göğüs kanseri tespiti yapmayı hedeflemekteydi. Diğer makine öğrenmesi algortimaları ile kıyaslandıktan sonra göğüs kanseri hastalığı tahmini için Destek Vektör Makinesi algoritmasının en iyi olduğu tespit edilmiştir.[6]
- Fabiano Texeria ve arkadaşları 2019 yılınd ayaptıkları çalışmada makine öğrenmesi algoritmalarının göğüs kanseri sınıflandırmadaki analizini incelemişlerdir. Değerlendirme için beş farklı sınıflandırma yöntemi kullanmışlardır: Çok Katmanlı Algılayıcı, Karar Ağacı, Rastgele Orman, Destek Vektör Makinesi ve Derin Sinir Ağı. Bu çalışma için, Wisconsin Üniversitesi Hastanesi veri tabanından faydalandıldı. DNN sınıflandırıcı doğruluk düzeyinde (%92) performansa sahiptir,bu da diğer modellere göre oldukça başarılıdır. Rastgele orman ROC eğrisi metriği için en iyi sonuçları sundu. [7]

II. MATERYAL VE YÖNTEM

Çalışmanın bu bölümünde veri seti hakkında bilgilendirme yapılmış olup kullanılacak yöntemlerin bilgisi verilmiştir.

Kullanılan veriseti, 1992 yılında Wisconsin Üniversitesi'nde yapılan bir çalışma sonucu toplanan verilerden oluşmaktadır. Bu veri seti kullanıma açık olup üzerinde bazı çalışmalar yapılmıştır. Veri seti 30 özellik ve bir hedef özellikten oluşmakta ve toplam 569 veriden meydana gelmektedir.

```
'mean radius', 'mean texture', 'mean perimeter', 'mean area',
'mean smoothness', 'mean compactness', 'mean concavity',
'mean concave points', 'mean symmetry', 'mean fractal dimension',
'radius error', 'texture error', 'perimeter error', 'area error',
'smoothness error', 'compactness error', 'concavity error',
'concave points error', 'symmetry error',
'fractal dimension error', 'worst radius', 'worst texture',
'worst perimeter', 'worst area', 'worst smoothness',
'worst compactness', 'worst concavity', 'worst concave points',
'worst symmetry', 'worst fractal dimension'], dtype='<U23')
```

Figure 1 Veri Seti Özellik İsimleri

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	...
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	...
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	...
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	...
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	...

Figure 2 Veri Seti Görüntüleri

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 31 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   mean radius              569 non-null    float64
 1   mean texture             569 non-null    float64
 2   mean perimeter           569 non-null    float64
 3   mean area                569 non-null    float64
 4   mean smoothness          569 non-null    float64
 5   mean compactness         569 non-null    float64
 6   mean concavity           569 non-null    float64
 7   mean concave points      569 non-null    float64
 8   mean symmetry            569 non-null    float64
 9   mean fractal dimension   569 non-null    float64
 10  radius error             569 non-null    float64
 11  texture error            569 non-null    float64
 12  perimeter error          569 non-null    float64
 13  area error               569 non-null    float64
 14  smoothness error         569 non-null    float64
 15  compactness error        569 non-null    float64
 16  concavity error          569 non-null    float64
 17  concave points error     569 non-null    float64
 18  symmetry error           569 non-null    float64
 19  fractal dimension error  569 non-null    float64
 20  worst radius             569 non-null    float64
 21  worst texture            569 non-null    float64
 22  worst perimeter          569 non-null    float64
 23  worst area               569 non-null    float64
 24  worst smoothness         569 non-null    float64
 25  worst compactness        569 non-null    float64
 26  worst concavity          569 non-null    float64
 27  worst concave points     569 non-null    float64
 28  worst symmetry           569 non-null    float64
 29  worst fractal dimension  569 non-null    float64
 30  Cancer                   569 non-null    int64
```

Figure 3 Veri Seti Ayrıntılı Gösterim

Çalışmada öncelikle veri seti hakkında bilgilendirmeler yapılmıştır. Özellikler hakkında kısaca bilgiler verilmiştir.

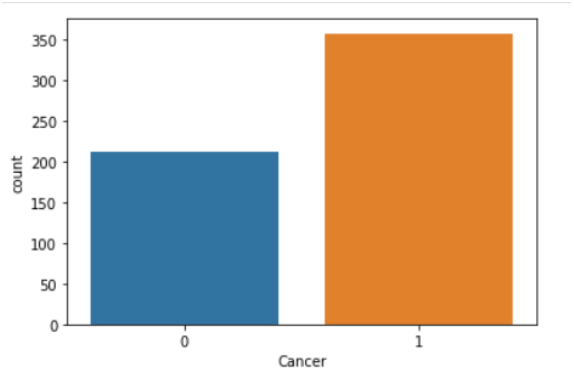

Figure 4 Kanserli ve Kanserli Olmayan Veri Gösterimi

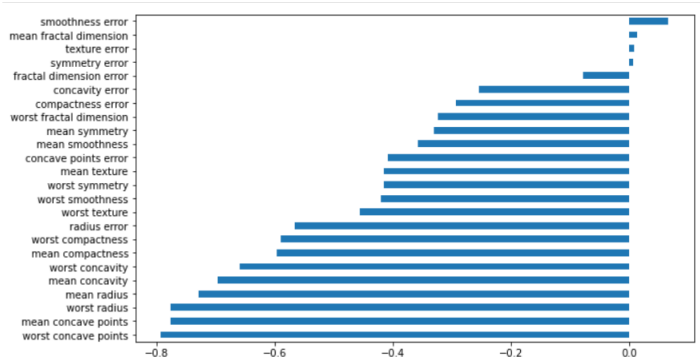

Figure 5 Özelliklerin Birbiri ile Korelasyonu

Veri setinde özellikler ve hedef tanımlamaları yapılmıştır. Bazı özellikler makine öğrenmesi kütüphaneleriyle birlikte görselleştirilmiştir. Bazı sütunların gereksiz olduğunu düşünülüp drop edilmiştir. Veriler arasındaki korelasyon kontrol edilmiştir.Ön işleme kısmında veriler test ve eğitim olarak rastgele şekilde ayrılmıştır. Bazı parametreler kullanılarak verinin okunması kolaylaştırılmıştır.

III. MODEL SEÇİMİ VE MODEL PERFORMANSI

Verinin boyutu da göz önüne alınarak aktivasyon fonksiyonları belirlendi ve uygulandı. Diğerlerine göre öğrenme hızı daha iyi olduğu için relu kullanıldı. Output değerler binary olduğu için sigmoid kullanıldı. Adam optimizeri kullanıldı. Sigmoid aktivasyon fonksiyonu aynı zamanda logistic fonksiyondu ve veriyi 0 ile 1 arasına sıkıştırıyordu. Çıkış katmanı için bu tercih edildi. Çıkış binary olduğundan, loss metriği olarak binarı_crossentropy kullanıldı. Eğitim için kullanılacak datanın %10 luk bölümü ayrıldı. O veri eğitime sokulmadı. Validasyon işlemi için kullanıldı. 1000 epochta eğitildi. loss ve val_loss iterasyonlar ilerledikçe aşağıya doğru azaldı. accuracy ve val_accurasy değeri de artıyor. 200 değerinden sonra overfit e gittiği tespit edildi. Modeli bu şekilde test ettiğimizde %91 sonucu alındı.

Table 1 Sonuç skorları

	Presicion	Recall	F1-score	Support
0	0.91	0.95	0.93	21
1	0.97	0.94	0.96	36
accurary			0.95	57
macro avg	*0.94*	*0.95*	*0.94*	*57*
weighted avg	0.95	0.95	0.95	57

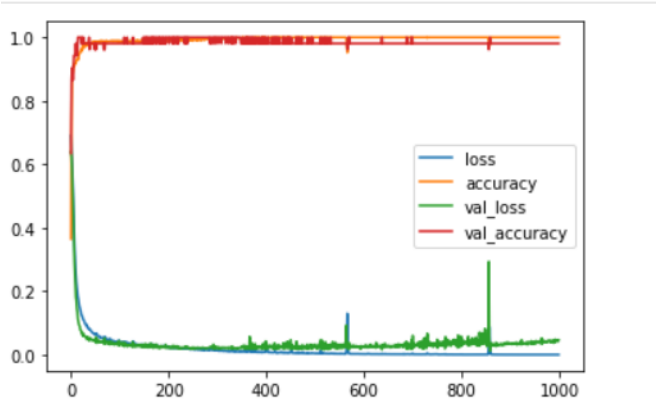

Figure 6 Loss ve Accurary Değerleri

A. *Earlystopping*

Aşırı öğrenmeyi önlemek için bazı yöntemler kullanıldı. Bunlardan birisi de earlystopping. EarlyStop kullanarak önceki aşırı öğrenmenin önüne geçildi. Earlystop val_loss değerini monitör edecek. patience değerini 15 alınarak daha iyi sonuçlar almak hedeflendi. Earlystop kullanarak overfittinge girmeden eğitimi durdurması sağlandı. Böylece sonuç %95 oldu.

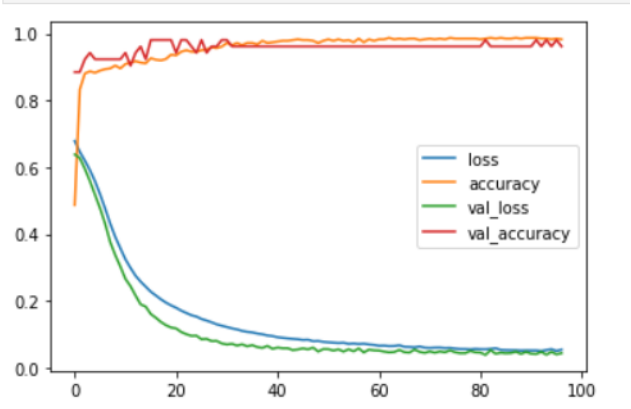

Figure 7 EarlyStopping değerleri

Earlystopp kullanarak aşırı öğrenmeyi durdurmuş olduk. 97. epoch'ta eğitim durdu.

Table 2 EarlyStop'tan sonraki sonuç skorları

	Presicion	Recall	F1-score	Support
0	1.00	0.95	0.98	21
1	0.97	1.00	0.99	36
accurary			0.98	57
macro avg	*0.99*	*0.98*	*0.98*	*57*
weighted avg	0.98	0.98	0.98	57

B. *Öğrenme Hızı*

Sonrasında Adam optimizeri ile birlikte learning_rate 0.005 olarak ayarlandı. Öğrenme hızı veya adım büyüklüğü dediğimiz parametre büyük olursa minimum olan noktaya ulaşılamaz, çok küçük olursa da model çok yavaş öğrenir. Learning rate ile adımlar büyütüldüğü için beklenildiği gibi skorda dalgalanma oldu.

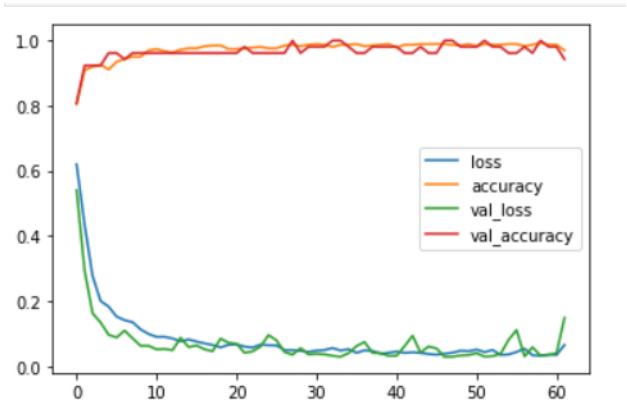

Figure 8 Öğrenme hızı ayarlandıktan sonraki loss ve accurary değerleri

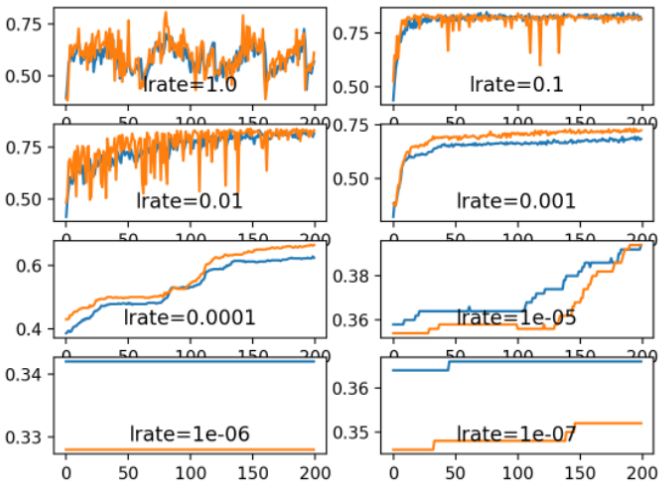

Figure 9 Öğrenme hızı etkisi

Table 3 Öğrenme hızı ayarlandıktan sonraki skorlar

	Presicion	Recall	F1-score	Support
0	1.00	0.86	0.92	21
1	0.92	1.00	0.96	36
accurary			0.95	57
macro avg	*0.96*	*0.93*	*0.94*	*57*
weighted avg	0.95	0.95	0.95	57

C. Dropout

Grafiğe bakıldığında, earlystop kullanmadan sadece Dropout ile overfitting düşürüldü. Lakin başarı oranı %93'te kaldı.

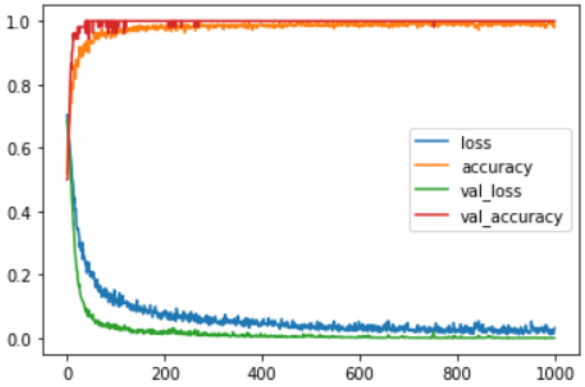

Figure 10 Droput Katmanı ile birlikte değerler

D. Çapraz Geçerleme

Çapraz geçerleme ile birlikte veri setinde bulunan herhangi bir grubu tüm veriyi alacak şekilde çaprazlayarak test edildi. Sonuçlar aşağıdaki tabloda mevcuttur.

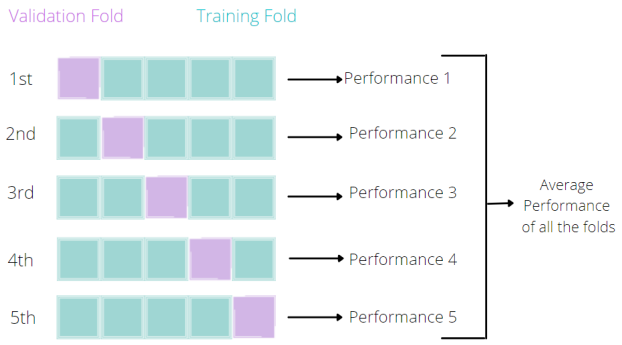

Figure 11 5 fold çapraz geçerleme

Table 4 Çapraz geçerleme ile oluşan sonuç skorları

	score_mean	score_std
acc	0.980468	0.024389
pre	0.979917	0.026112
rec	0.990892	0.020197
f1	0.985170	0.018417

IV. DENEYSEL SONUÇLAR

A. GridSearchCV

Bazı katmanlar ekleyip GridSearchCV metodu kullanılarak iyileştirilmeler yapılmış ve başarı oranı %97'ye kadar çıkarılmıştır. ROC (Receiver Operating Curve) and AUC (Area Under Curve) for grid_model eğrileri aşağıdaki görselde mevcuttur.

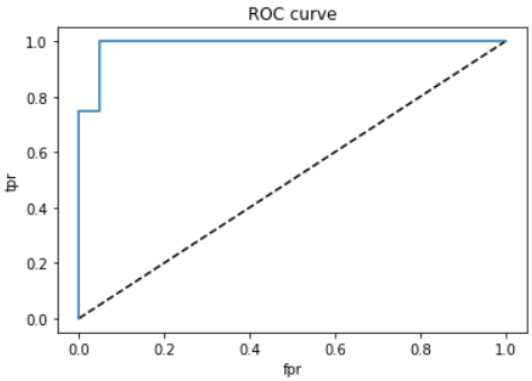

Figure 12 GridSearchCV için ROC Eğrisi

Sonrasında ise modeli kaydedip tahmin işlemi yapılması gerçeklenmiştir.

B. Lojistik Regresyon

Table 5 Lojistik Regresyon Skorları

	precision	recall	f1-score	support
0	0.95	0.95	0.95	21
1	0.97	0.97	0.97	36
accuracy			0.96	57
macro avg	0.96	0.96	0.96	57
weighted avg	0.96	0.96	0.96	57

C. Rastgele Orman

Table 6 Rastgele Orman Skorları

	precision	**recall**	**f1-score**	**support**
0	0.95	0.95	0.95	21
1	0.97	0.97	0.97	36
accuracy			0.96	57
macro avg	0.96	0.96	0.96	57
weighted avg	0.96	0.96	0.96	57

V. CONCLUSION

Makine öğrenmesi algoritmalarından GridSearchCV ile %97Logistic Resresyon ile %96, Rastgele Orman ile %96 sonuçları bulunmuştur. Çeşitli parametlerler iyileştirmeler yapıldı ve en iyi sonuç olarak GridSearchCV olarak bulundu. Göğüs Kanseri veri setinde kanserin erken teşhisi için makine öğrenmesi tekniklerinin önemine binaen geliştririlen bu yöntem ile sağlık sektörüne katkıda bulunulması hedeflenmektedydi. Bu veri seti ile çalışan bir çok araştırmacı var ve her biri farklı metotlar denemiş ve farklı sonuçlar elde etmiştir. Bizim yöntemlerimizde bazı parametreler göz önüne alınarak GridSearchCv metodu en yüksek başarı oranına sahip metot olarak bulunmuştur.

REFERENCES

[1] A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," in *2010 5th international symposium on health informatics and bioinformatics*. IEEE, 2010, pp. 114–120.

[2] A. Qasem, S. N. H. S. Abdullah, S. Sahran, T. S. M. T. Wook, R. I. Hussain, N. Abdullah, and F. Ismail, "Breast cancer mass localization based on machine learning," in *2014 IEEE 10th International Colloquium on Signal Processing and its Applications*. IEEE, 2014, pp. 31–36.

[3] N. Khuriwal and N. Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," in *2018 IEEMA Engineer Infinite Conference (eTechNxT)*, 2018, pp. 1–5.

[4] E. A. Bayrak, P. Kırcı, and T. Ensari, "Comparison of machine learning methods for breast cancer diagnosis," in *2019 Scientific Meeting on Electrical-Electronics Biomedical Engineering and Computer Science (EBBT)*, 2019, pp. 1–3.

[5] S. Ara, A. Das, and A. Dey, "Malignant and benign breast cancer classification using machine learning algorithms," in *2021 International Conference on Artificial Intelligence (ICAI)*, 2021, pp. 97–101.

[6] A. Mangal and V. Jain, "Prediction of breast cancer using machine learning algorithms," in *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2021, pp. 464–466.

[7] F. Teixeira, J. L. Z. Montenegro, C. A. da Costa, and R. da Rosa Righi, "An analysis of machine learning classifiers in breast cancer diagnosis," in *2019 XLV Latin American Computing Conference (CLEI)*, 2019, pp. 1–10.