Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

GU J Sci, Part C, 12(2): 675-690 (2024)

Öznitelik Seçimi ile Desteklenen Makine Öğrenmesine Dayalı Göğüs Kanserinin Erken Tespiti ve Teşhisi

Cihan AKYEL[1*], Bünyamin CİYLAN [2], Hüseyin POLAT [2]

[1]*Millî Eğitim Bakanlığı, İnşaat ve Emlak Genel Müdürlüğü, Beşevler Kampüsü, Ankara, Turkey*

[2]*Gazi University, Teknoloji Fakültesi, Bilgisayar Mühendisliği Bölümü, Ankara, Turkey*

Makale Bilgisi

Araştırma makalesi
Başvuru: 07/12/2023
Düzeltme: 13/02/2024
Kabul: 01/03/2024

Anahtar Kelimeler

Göğüs Kanseri
Öznitelik Seçimi
Makine Öğrenmesi
Karar Destek Sistemleri
Bilişim Sistemleri
Veri Yönetimi

Article Info

Research article
Received: 07/12/2023
Revision: 13/02/2024
Accepted: 01/03/2024

Keywords

Breast Cancer
Feature Selection
Machine Learning
Decision Support Systems
Information Systems
Data Management

Grafik Özet (Graphical/Tabular Abstract)

Bu çalışmada göğüs kanserinin tespiti için Wisconsin veri setinde öznitelik seçimi yöntemleri ile öznitelik sayısı beşe indirilerek "iyi huylu" ve "kötü huylu" olarak sınıflandırılmıştır. Farklı öznitelik seçimi ve makine öğrenmesi yöntemleri denenerek en yüksek başarıyı gösteren ikili seçilmiştir. Teşhis sürecinde karar desteği sağlamak adına bir karar destek sistemi oluşturulmuştur. / In this study, for the detection of breast cancer, the number of features in the Wisconsin data set was reduced to five using feature selection methods and classified as "benign" and "malignant". By trying different feature selection and machine learning methods, the pair with the highest success was selected. A decision support system was created to provide decision support in the diagnosis process of breast cancer.

Şekil A: *Çalışmaya ait genel akış /* ***Figure A:*** *General flow of the study*

Önemli noktalar (Highlights)

- *Farklı öznitelik seçimi ve makine öğrenmesi yöntemleri denenerek en uygun ikili seçildi. / Different feature selection and machine learning methods were tried and the most suitable pair was selected.*
- *Çalışmada göğüs kanserinin teşhisi süreci için karar destek sistemi tasarlandı. / In the study, a decision support system was designed for the breast cancer diagnosis process.*
- *Daha az öznitelik kullanılarak daha az kaynak ile yüksek başarı elde edildi. / High success was achieved with fewer resources by using fewer attributes.*

Amaç (Aim): *Çalışmada bellek, gpu gücü gibi parametrelerde daha az kaynak kullanılarak, göğüs kanserinin ikili sınıflandırılması ve bir karar destek sistemi üzerinden teşhis sürecine destek sağlanması amaçlanmaktadır. / The aim of the study is to dual-classify breast cancer and provide support to the diagnosis process through a decision support system, by using fewer resources in parameters such as memory and gpu power.*

Özgünlük (Originality): *Çalışmada farklı öznitelik seçimi ve makine öğrenmesi yöntemlerine ait olası tüm kombinasyonlar için eğitimler yapılmıştır. Hangi özniteliklerin teşhis için daha etkili olduğu ortaya konulmuştur. Daha az öznitelik ile uygun öznitelik seçimi ve makine öğrenmesi yöntemleri kullanılarak da yüksek başarı elde edilebileceği görülmüştür. Ayrıca çalışma kapsamında karar destek sistemi oluşturularak bu alanda farklılık sağlanmıştır. / In the study, training was conducted for all possible combinations of different feature selection and machine learning methods. It has been revealed which features are more effective for diagnosis. It has been observed that high success can be achieved by using appropriate feature selection and machine learning methods with fewer features. In addition, a decision support system was created within the scope of the study and a difference was made in this field.*

Bulgular (Results): *%98,83 doğruluk, %99 kesinlik ve %99 duyarlılık değerleri, Variance inflation factors (VIF) öznitelik seçimi ve Random Forest algoritması kullanılarak elde edilmiştir. / It was obtained using Variance inflation factors (VIF) feature selection and Random Forest algorithm with 98.83% accuracy, 99% precision and 99% sensitivity values.*

Sonuç (Conclusion): *Bu çalışma ile kanserin erken teşhisinde karar vericilere yardımcı olacak bir sistem sunulmaktadır. / This study presents a system that will assist decision makers in the early diagnosis of cancer.*

*Corresponding author, e-mail: cihan.akyel06@gmail.edu.tr
DOI: 10.29109/gujsc.1400991

GU J Sci, Part C, 12(2): 675-690 (2024)

Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

http://dergipark.gov.tr/gujsc

Öznitelik Seçimi ile Desteklenen Makine Öğrenmesine Dayalı Göğüs Kanserinin Erken Tespiti ve Teşhisi

Cihan AKYEL[1*], Bünyamin CİYLAN[2], Hüseyin POLAT[2]

[1]*Millî Eğitim Bakanlığı, İnşaat ve Emlak Genel Müdürlüğü, Beşevler Kampüsü, Ankara, Turkey*

[2]*Gazi University, Teknoloji Fakültesi, Bilgisayar Mühendisliği Bölümü, Ankara, Turkey*

Makale Bilgisi

Araştırma makalesi
Başvuru: 07/12/2023
Düzeltme: 13/02/2024
Kabul: 01/03/2024

Anahtar Kelimeler

Göğüs Kanseri
Öznitelik Seçimi
Makine Öğrenmesi
Karar Destek Sistemleri
Bilişim Sistemleri
Veri Yönetimi

Öz

Kanserin tam nedeni bilinmemekle birlikte, yaşam tarzı, çevresel faktörler, beslenme ve genetik gibi birçok faktörün kanser gelişimine katkıda bulunabileceği bilinmektedir. Kanser türleri arasında özellikle göğüs kanseri, dünya genelinde kadınlar arasında görülme sıklığı yüksek olan bir hastalıktır. Göğüs kanserinin teşhisinde fiziksel muayene ve mamografi görüntülerinin incelenmesi gibi yöntemler kullanılmaktadır. Gelişen teknolojiyle birlikte makine öğrenmesi uygulamalarının tıp alanında kullanımı giderek artmaktadır. Bu sayede göğüs kanserinin daha erken aşamada ve hızlı şekilde teşhisi konusunda doktorlara yardımcı olabilecek umut verici çalışmalar giderek artmaktadır. Bu çalışmada, göğüs kanserinin erken teşhisinde kullanmak için 4 farklı öznitelik seçimi ve 5 farklı makine öğrenme yönteminin performansları karşılaştırılmıştır. Çalışmanın ilk aşamasında, Principal Component Analysis (PCA), Recursive feature elimination, Variance inflation factors (VIF) ve Univariate feature selection yöntemleri ile veri kümesinde hedef özniteliğe en çok etki eden öznitelikler seçilerek veri kümesindeki öznitelik sayısı azaltılmıştır. İkinci aşamada, K Nearest Neighbors (KNN), Naive Bayes, Decision Tree, Support Vector Machine (SVM) ve Random Forest makine öğrenme algoritmaları orijinal ve öznitelik seçimi yapılmış veri kümelerine dayalı olarak eğitilmiş ve test edilmiştir. Test sonuçlarına göre %98,83 doğruluk, %99 kesinlik ve %99 duyarlılık değerleri ile Variance inflation factors (VIF) öznitelik seçimi ve Random Forest algoritması kullanılarak elde edilmiştir. Daha az öznitelik kullanımı sayesinde eğitim ve test aşamalarında benzer başarı değerleri, kaynak kullanımı ile sağlanmıştır. Çalışmada eğitilip test edilen makine öğrenme modeli Flask framework kullanılarak bir web ara yüzüne sahip uygulama haline getirilmiştir.

Early Detection and Diagnosis of Breast Cancer Based on Machine Learning Supported by Feature Selection

Article Info

Research article
Received: 07/12/2023
Revision: 13/02/2024
Accepted: 01/03/2024

Keywords

Breast Cancer
Feature Selection
Machine Learning
Decision Support Systems
Information Systems
Data Management

Abstract

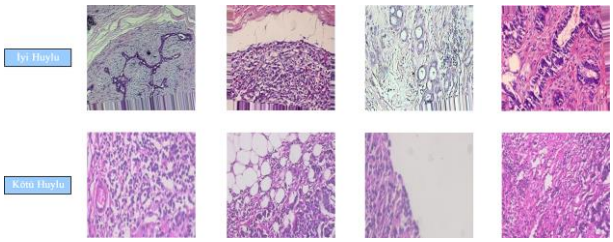
Although the exact cause of cancer is unknown, it is known that many factors such as lifestyle, environmental factors, nutrition and genetics can contribute to the development of cancer. Among the types of cancer, especially breast cancer is a disease with a high incidence among women worldwide. Methods such as physical examination and examination of mammography images are used to diagnose breast cancer. With developing technology, the use of machine learning applications in the medical field is increasing. In this way, promising studies that can help doctors diagnose breast cancer at an earlier and faster stage are increasing. In this study, the performances of 4 different feature selection and 5 different machine learning methods were compared for use in the early diagnosis of breast cancer. In the first stage of the study, the number of features in the dataset was reduced by selecting the features that most affect the target feature in the dataset using Principal Component Analysis (PCA), Recursive feature elimination, Variance inflation factors (VIF) and Univariate feature selection methods. In the second stage, K Nearest Neighbors (KNN), Naive Bayes, Decision Tree, Support Vector Machine (SVM) and Random Forest machine learning algorithms were trained and tested based on original and feature selected datasets. According to the test results, 98.83% accuracy, 99% precision and 99% sensitivity values were obtained using Variance inflation factors (VIF) feature selection and Random Forest algorithm. Thanks to the use of fewer features, similar success values were achieved in the training and testing phases with the use of resources. The machine learning model trained and tested in the study was turned into an application with a web interface using the Flask framework.

*Corresponding author, e-mail: cihan.akyel06@gmail.com
DOI: 10.29109/gujsc.1400991

1. GİRİŞ (INTRODUCTION)

Kanser, yaşam tarzı, beslenme alışkanlıkları, genetik etmenler ve çevresel faktörlere bağlı olarak gelişebilen ve dünya genelinde görülen sağlık sorunları arasında büyük öneme sahip bir hastalıktır. Her yıl milyonlarca insan kanser nedeniyle yaşamını yitirmektedir. Kanserin etkileri sadece bireylerle sınırlı kalmamakta, aynı zamanda ekonomi ve sağlık sistemi üzerinde de büyük bir yük oluşturmaktadır. Kanser türlerinin kadınlar ve erkekler arasındaki görülme ihtimali ve sıklığı değişkenlik gösterebilmektedir. Göğüs kanseri özellikle kadınlarda, dünya genelinde en yaygın görülen kanser türüdür. Göğüs kanseri, meme dokusunda anormal hücrelerin kontrolsüz şekilde çoğalıp tümör oluşturması ve çevre dokulara yayılması ile ortaya çıkmaktadır. Bununla birlikte bir tümör her durumda kanser anlamına da gelmemektedir. Tümörler iyi huylu eğilim de gösterebilmektedirler. İyi huylu tümörler vücudun diğer bölgelerine yayılmadıkları için kanser olmayan tümörler olarak kabul edilirler. Kötü huylu tümörler ise kanser olarak adlandırılırlar ve çoğalarak vücudun diğer bölgelerine yayılabilirler. Kötü huylu kanser vakaları hastalar için ölümle sonuçlanan durumlar oluşturabilmektedirler. Dünya Sağlık Örgütü'nün verilerine göre 2020 yılında yaklaşık 2,3 milyon yeni göğüs kanseri vakası görülmüş olup, bu vakaların 685.000'i ölümle sonuçlanmıştır [1]. Bu veriler, göğüs kanserini 2020 yılında en ölümcül beşinci kanser türü yapmaktadır. Göğüs kanserinde erken teşhis ve tedavi gerçekleşmezse kanser hücreleri lenfatik sistem aracılığıyla vücudun diğer bölgelerine yayılabilmektedir. Bu durum da tedavi sürecini hasta ve sağlık sistemi açısından daha zorlu hale getirmektedir. Bu yüzden göğüs kanserinin erken teşhisi hastaların tedavi başarısını artırması açısından son derece önemlidir. Erken teşhis durumunda, göğüs kanseri vakalarında hayatta kalma oranları %97'lere kadar çıkabilmektedir [2].

Göğüs kanserinin teşhis sürecinde mamografi, ultrason, manyetik rezonans görüntüleme, ince iğne aspirasyonu (fine needle aspirate) ve cerrahi biyopsi gibi yöntemlerden faydalanılır [3]. Doğru bir teşhis için bu yöntemlerin duyarlılık sonuçları şu şekilde değişmektedir: mamografi %68-%79, ince iğne aspirasyonu %65 - %98 ve cerrahi biyopsi yaklaşık %100 [4]. Kesin teşhis için cerrahi biyopsi etkili bir yöntem olmasına rağmen, hastalar üzerinde olumsuz psikolojik davranışlara neden olan maliyetli bir prosedür olarak bilinmektedir. Göğüs kanserinin teşhisinde kullanılan bir diğer etkili yöntem de, tümörden hücre örnekleri alma ve mikroskop altında hücresel gözlem yapma tekniği olan ince iğne aspirasyon biyopsisidir. İnce iğne aspirasyonunun birçok avantajı ve potansiyel kullanımı vardır, bu da onu göğüs kanserinin tanısı için popüler bir seçenek haline getirmektedir. İnce iğne aspirasyonu, meme lezyonlarından doku ve sıvı örnekleri almak için küçük (21 ile 25 kalibrelik) bir iğne kullanılarak yapılan bir biyopsi türüdür. Doktorlar tümör veya tümörlerin derinliğini, sayısını ve yerini belirlemek için ince iğne aspirasyonundan önce bir ultrason isteyebilirler. Kanser hücrelerinin özelliklerinin bilinmesi, kanserin teşhisi için büyük önem taşır. Bir hastadaki kanserli hücrelerin belirlenmesi oldukça özneldir ve hekim uzmanlığına bağlıdır. Bu durum göğüs kanseri hücrelerinin tespiti ve sınıflandırılması için doğru araçlara yönelmeyi gerektirmektedir. İnce iğne aspirasyonu yardımıyla göğüs kanseri teşhis sürecinin doğruluğunu, tutarlılığını ve verimliliğini artırmak için, hücre görüntülerinden hücre çekirdeklerinin şekil bilgisi ile teşhis yapılmaya çalışılır [5]. Günümüzde bu işlemi bir örüntü sınıflandırma problemi olarak ele alan makine öğrenme teknikleri kullanılmaktadır. İyi huylu ve kötü huylu lezyonlara ait örnekler Şekil 1'de görülmektedir.



Şekil 1. İyi huylu ve kötü huylu lezyon örnekleri (Examples of benign and malign lesions)

Teknolojik gelişmeler sayesinde makine öğrenme yöntemleri, göğüs kanserinin erken teşhisinde büyük bir potansiyel taşımaktadır ve makine öğrenmesi teşhis sürecini daha hassas, etkili ve hızlı hale getirebilir. Göğüs kanseri teşhisinde hücre çekirdeği ölçümleri, kanser hücrelerinin içsel özelliklerini anlamada kritik bir rol oynar. Kanser hücreleri genellikle büyük, anormal şekilli ve hızlı bölünen çekirdeklere sahip olabilir. Bu hücre çekirdek özellikleri, makine öğrenmesi algoritmaları tarafından değerlendirilebilir ve kanserli hücreleri tespit etmek için kullanılabilir. Bu sayede, kanserli hücreler daha hızlı ve hassas şekilde belirlenebilir.

Makine öğrenmesi genel olarak verilerin toplanması, makine öğrenme modelinin eğitimi ve test aşamalarını içeren bir süreçtir. Makine öğrenmesi yöntemleriyle hastalık teşhisinde eğitim

ve test süreçlerinde veri kümeleri önemli bir yer tutmaktadır. Vakalara ait verilerin yer aldığı bu veri kümelerinde birçok öznitelik yer alabilmektedir. Bu öznitelikler hedef (teşhis) ile değişen oranlarda korelasyon göstermektedirler. Öznitelik seçim yöntemleri ile hedef ile korelasyonu yüksek öznitelikler seçilerek veri kümesinin boyutu azaltılabilmektedir. Öznitelik seçimi, sınıflama çalışmalarında alakasız öznitelikleri veri kümesinden kaldırarak doğruluk değerini artırır [6]. Veri kümelerinin ham halleriyle eğitime alınması eğitim süresini ve modelin cevap süresini artırmaktadır. Bu çalışmada "Breast Cancer Wisconsin (Diagnostic) Data Set" adlı veri kümesi üzerinde öznitelik seçimi teknikleri uygulanmıştır. Daha sonra farklı makine öğrenme teknikleriyle sınıflandırma yapılmıştır. Deneysel çalışmalar sonucunda en yüksek başarıyı gösteren öznitelik seçimi ve makine öğrenme yöntemi seçilmiştir. Ayrıca bu yöntem ikilisinin uygulama olarak kullanılabilmesi için bir web ara yüzüne sahip bir karar destek sistemi geliştirilmiştir. Oluşturulan karar destek sistemi ile veri yönetimi kapsamında veriler csv uzantılı bir dosyada saklanarak karar vericilere sunulan sonuçlar için işlenmektedir [7].

1.1. Literatür Taraması (Literature Review)

Günümüzde hemen her alanda olduğu gibi sağlık alanında da veri yönetimi karar süreçleri açısından önemlidir. Verilerin işlenmesi ve sınıflandırılması veri yönetiminin önemli aşamalarındandır. Veri yönetimiyle ilgili yapılan bir çalışmada sağlıkta veri kalitesi ve veri madenciliği uygulamaları araştırılmıştır [8]. Ultrason görüntülerini (OASBUD veri seti) kullanarak yapılan bir çalışmada MobileNetV2 modeli ile %71 başarı oranı elde edilmiştir [9]. Doğan vd. sundukları çalışmada göğüs kanserinin makine öğrenme teknikleriyle teşhisi konusu yer almıştır. Klinik sayısal bulguların yer aldığı bir veri kümesi kullanılmıştır. Bu çalışmada farklı yöntemler denenmiş olup K en yakın komşu algoritmasının en yüksek başarıyı gösterdiği belirtilmiştir K en yakın komşu algoritması ile yapılan sınıflandırmada %99,42 doğruluk oranı elde edilmiştir [10]. Bozkurt Keser ve Keskin, tarafından sunulan çalışmada göğüs kanseri sınıflandırılmıştır. Çalışmada Wisconsin Üniversitesine ait görüntülerin yarıçap, doku ve çevre gibi farklı öznitelikleri bulunduran bir veri kümesi kullanılmıştır. Önerilen sınıflama yöntemi ile %98,77 doğruluk oranı gözlemlenmiştir [11].

Erdem ve Aydın, histopatolojik görüntülerin sınıflandırılması ilgili bir çalışma sunmuşlardır. Çalışmada iki farklı model olan Inception-V3 ve VGG16 birlikte kullanılarak hibrit bir yaklaşım ortaya konmuştur. Veri kümesi olarak göğüs kanserine ait histopatolojik görüntüleri içeren BreakHis adlı toplam 9109 görüntü içeren bir veri kümesi tercih edilmiştir. Çalışma sonucunda önerilen yaklaşım ile %98,3 doğruluk oranı elde edilmiştir. Veri kümesi üzerinde veri artırma ve yeniden boyutlandırma işlemleri gerçekleştirilmiştir [12].

ResNet esnek yapısından dolayı birçok farklı türe sahip olan ve görüntü sınıflandırma çalışmalarında sıklıkla tercih edilen bir modeldir. Bu çalışmalardan biri Talo tarafından sunulmuştur. Çalışmada bir ResNet türevi olan 50 katmana sahip ResNet-50 modeli eğitim aşamalarında tercih edilmiştir. Çalışmada BreakHis veri seti kullanılmıştır. Bu veri ile ResNet-50 modeli kullanılarak 40X büyütme oranına sahip veriler ile %98,83 doğruluk oranı elde edilmiştir [13].

Görüntülerin sınıflandırılmasında ön eğitimli modellerin kullanılması başarı artışını sağlayan, aynı zamanda eğitim süresini azaltabilen bir yaklaşımdır. Spanhol vd. tarafından sunulan çalışmada ön eğitimli kullanıma sahip olan AlexNet temelli bir algoritma kullanılmıştır. BreakHis veri setinin kullanıldığı bu çalışmada %85,6 doğruluk oranı gözlemlenmiştir [14]. Han vd. tarafından sunulan çalışmada BreakHis veri kümesi veri artırma yöntemleri ile birlikte kullanılmıştır. Bu çalışmada veriler üzerinde normalizasyon uygulanmıştır. Çalışmada CSDCNN adlı bir derin öğrenme algoritması önerilmiştir. Bu model kullanılarak %96,90 doğruluk oranı 100x büyütülmüş verilerle elde etmişlerdir [15].

IRRCNN adlı derin öğrenme modeli kullanılarak yapılan çalışmada veri artırma yöntemlerinin de desteğiyle %97,95 doğruluk oranı elde edilmiştir. BreakHis veri kümesi bu çalışmada tercih edilerek eğitim aşaması tamamlanmıştır. Burada paylaşılan başarı değeri veri kümesinde yer alan 40X büyütülen görüntülerle yapılan eğitim sonucunda elde edilen başarı sonucudur [16]. Kahya vd. histopatolojik görüntülerle göğüs kanserinin teşhisi için bir çalışma sunmuşlardır. BreakHis veri kümesi halka açık olması ve farklı büyütme oranlarından dolayı birçok çalışmada tercih edilmektedir. Destek vektör makineleri algoritması kullanılarak en yüksek başarı olan %96,28, 200x büyütme

oranındaki verilerle yakalanmıştır. Çalışmada aktarılan ASSVM modeli ise en yüksek başarıyı 40x büyütme oranı kullanıldığında %94,97 ile elde etmiştir [17]. Lavanya ve Rani tarafından sunulan çalışmada CART algoritması tercih edilmiştir. Bu çalışmada veri seti olarak Breast Cancer Wisconsin (Diagnostic) kullanılmıştır. %92,27 başarı değeri orijinal veri seti ile elde edilmiştir. PCVA öznitelik seçimi yöntemi ile öznitelik sayısı 9'a düşürülerek yeni veri seti elde edilmiştir. Bu veri seti ile %96,99 doğruluk oranı elde edilmiştir. Bu çalışma öznitelik seçimi yöntemleri ile daha az özniteliğe sahip veri setleri elde edilebildiği görülmektedir. Ayrıca bu durumun başarıyı artırabildiği de ortaya konulmuştur [6].

BreakHis veri setindeki görüntüler 40X, 100X, 200X ve 400X olmak üzere dört farklı şekilde bulunmaktadır. Gupta vd. tarafından sunulan çalışmada BreakHis veri seti kullanılmıştır. En yüksek başarı 200X büyütülmüş görüntülerle %88,89 doğruluk oranı şeklinde elde edilmiştir. Farklı yöntemler birleştirilerek kullanılmıştır [18]. Dandıl ve Serin sundukları çalışmada BreakHis veri setini Xception modeli ile eğitimde kullanmışlardır. Çalışmada %98,11 doğruluk oranı, %97,89 kesinlik (precision) ve %97,47 duyarlılık (recall) değerleri elde edilmiştir [19]. Narin ve Kefeli tarafından sunulan çalışmada ResNet ve VGG16 algoritmalarını kullanmışlardır. Ön eğitimli ResNet50 ile 200X görüntüler kullanılarak en yüksek başarısı olan %93,03 doğruluk oranı elde edilmiştir (hassaslık-sensitivity %92,81, özgüllük-specificity %93,55). VGG16 ile en yüksek hassasiyet değeri %99,28 olarak gözlemlenmiştir (doğruluk oranı %93,03, özgüllük-specificity %79,03) [20].

Sağlık alanında karar vericilere yardımcı olacak karar destek sistemlerinin kullanım örnekleri literatürde yer almaktadır. Ülkemizde kullanılan Merkezi Hekim Randevu Sistemi (MHRS) sistemi de karar destek sistemi olarak görülebilmektedir [21]. Bu alanda yapılan diğer bir çalışmada covid-19 teşhisi için bir karar destek sistemi oluşturulmuştur. Model bileşeninde yapay zeka kullanılan bu karar destek sistemi ile %90 doğruluk değeri elde edilmiştir [22]. Geliştirilen diğer bir karar destek sistemi ile hemşirelerin yapacakları hasta risk değerlendirmelerinde harcayacakları sürenin azaltılması sağlanmıştır. Farklı durumlar için oluşturulan algoritmalar karar destek sisteminin model bileşenini oluşturmuştur [23].

Göğüs kanserinin teşhisinde literatürde genellikle histopatolojik görüntüler kullanılarak yapılan çalışmalar vardır. Bunun yanı sıra "breast cancer Wisconsin" gibi görüntülere ait öznitelik bilgilerini barındıran veri kümelerinin kullanıldığı çalışmalar da mevcuttur. Bu çalışmalardan birisi Mohammed vd. tarafından sunulmuştur. Çalışmada K en yakın komşu (K-nearest neighbor-KNN) ve destek vektör makinası (Support vector machine-SVM) gibi farklı modeller ile eğitimler yapılmıştır. En yüksek doğruluk oranı SVM ile %97,70 olarak elde edilmiştir [24]. Bu veri kümesinin kullanıldığı diğer bir çalışmada SVM, KNN ve random forest gibi modeller ile eğitim yapılmıştır. Çalışma sonucunda en yüksek doğruluk oranı random forest yöntemi ile %98,77 olarak elde edilmiştir. SVM yönteminin kullanıldığı bir diğer çalışmada quadratic SVM ile göğüs kanserinin sınıflandırılması yapılmıştır. Breast Cancer Wisconsin (Diagnostic) veri setinin kullanıldığı çalışmada %98,1 doğruluk oranı elde edilmiştir [25]. Agarap tarafından yapılan bir çalışmada SVM yöntemi ve Breast Cancer Wisconsin (Diagnostic) Data Set veri seti kullanılarak %96,09 doğruluk değeri elde edilmiştir [26].

Çalışmanın ikinci bölümde kullanılan veri kümesi ve yöntemler detaylandırılmıştır. Üçüncü bölümde elde edilen sonuçlar yöntemlere göre karşılaştırılarak sunulmuştur. Dördüncü bölümde geliştirilen karar destek sistemi örnek değerler üzerinden açıklanmıştır. Beşinci bölümde ise çalışmada elde edilen sonuçlar ele alınarak literatüre olan katkısı üzerinde durulmuştur.

2. MATERIALS AND METHODS (MATERYAL VE METOD)

Bu çalışmada, öncelikle mevcut veri kümesi üzerinde Principal Component Analysis (PCA), Variance inflation factors (VIF), Recursive feature elimination ve Univariate feature selection öznitelik seçim yöntemleri uygulanarak veri kümesinin yeni versiyonları oluşturulmuştur. Bu yöntemler temelde bir dizi muhtemelen ilişkili değişkeni, temel bileşenler adı verilen daha az sayıda değişkene dönüştürmeyi hedefleyen matematiksel ilkeleri kullanmaktadırlar [27, 28]. İkinci aşamada, orijinal ve öznitelik seçimi uygulanmış veri kümeleri üzerinde farklı makine öğrenme yöntemleri ile eğitim ve test işlemleri gerçekleştirilmiştir. Makine öğrenme yöntemleri olarak literatürde yerini almış ve birçok sınıflandırma çalışmasında tercih edilen toplamda 5 farklı algoritma seçilerek kullanılmıştır.

Bunlar KNN (K Nearest Neighbors), Naive Bayes, Decision Tree, SVM ve Random Forest algoritmalarıdır. Her bir yöntem hem orijinal veri kümesiyle hem de öznitelik seçimi uygulanmış veri kümeleriyle eğitilmiş ve test edilmiştir.

2.1. Veri Kümesi (Dataset)

Bu çalışmada "Breast Cancer Wisconsin (Diagnostic) Data Set" adlı veri kümesi kullanılmıştır. Veri kümesi, University of Wisconsin-Madison Hastanesi'nde 1995 yılında toplanan ve Dr. William H. Wolberg tarafından oluşturulan bir veritabanına dayanmaktadır. Veri kümesinde, hücreler hakkında çeşitli bilgiler içeren toplam 30 farklı öznitelik (id ve diagnosis alanları hariç) bulunmaktadır. Öznitelikler, bir meme kitlesinin ince iğne aspirasyonunun sayısallaştırılmış görüntüsünden hesaplanır. Bu öznitelikler, kanser hücrelerine ait lezyon görüntülerinin hücre çekirdeği ölçümlerini ve diğer biyolojik özellikleri temsil eder. Bu veri kümesinde 357 iyi huylu ve 212 kötü huylu olmak üzere toplam 569 örnek yer almaktadır. Tüm veri kümesinin %62,7'sine tekabül eden 357 örnek iyi huylu hücrenin varlığını, %37,3'üne tekabül eden 212 örnek ise kanserli hücrenin varlığını göstermektedir [29].

Her hücre çekirdeği için on (10) gerçek değerli ana öznitelik bulunmaktadır (Tablo 1). Her bir ana öznitelik için ortalama (mean), standart hata (se) ve en büyük 3 değerin ortalaması (worst) şeklinde 3 öznitelik değeri (radius_mean, radius_se ve radius_worst gibi) hesaplanmıştır (30 öznitelik) [29].

Table 1. Veri kümesinde yer alan temel öznitelikler (Key attributes included in the dataset)

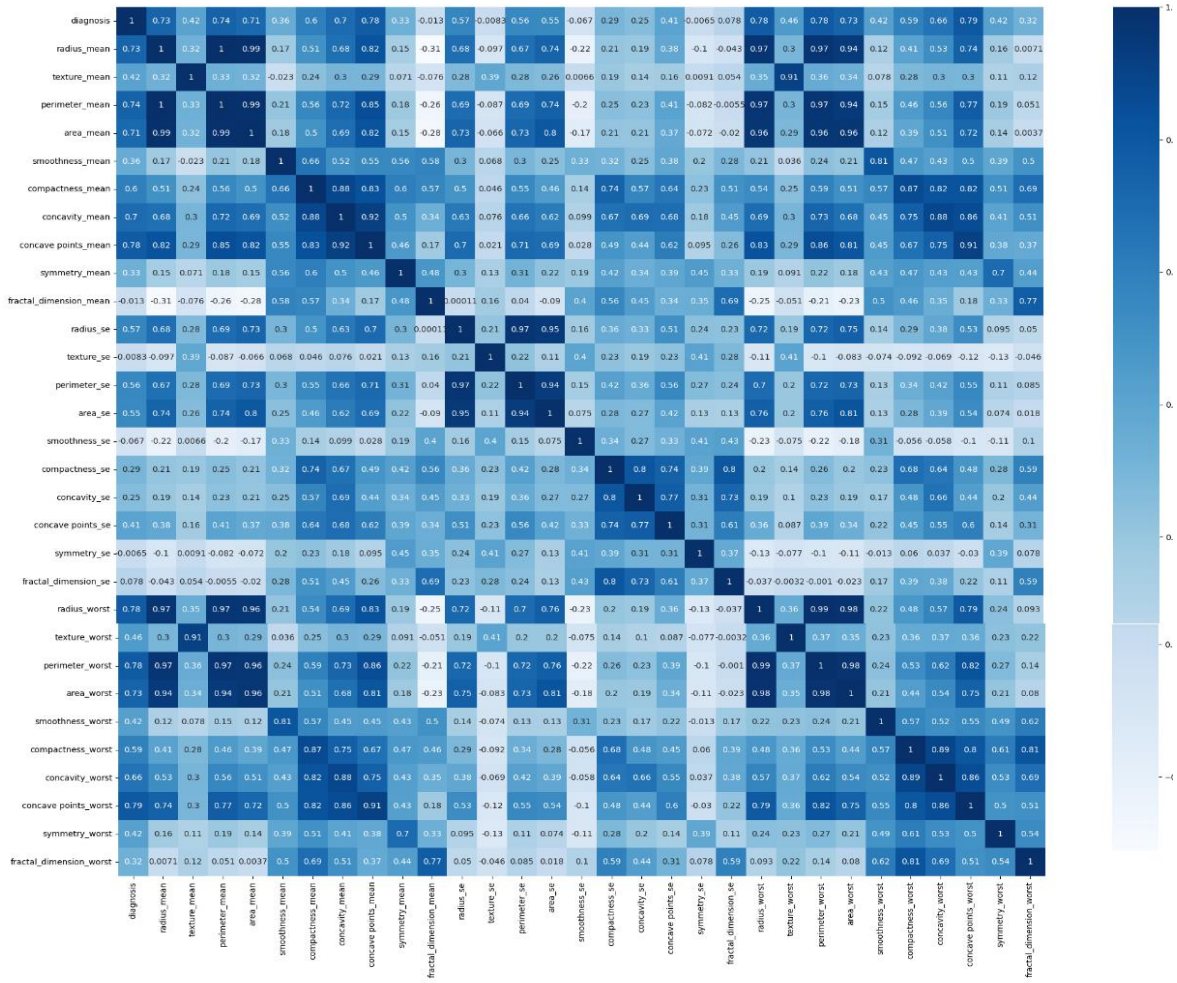
	Temel Öznitelik	**Açıklama**	**Öznitelikler**
1	Radius	Hücrelerin yarıçap değerlerinin ortalama (mean), standart hata (se) ve en kötü değeri (worst) hesaplanır.	radius_mean radius_se radius_worst
2	Texture	Hücrelerin iç yüzeylerinin gri tonlama değişim oranlarının ortalaması (mean), standart hatası (se) ve en kötü değeri (worst) hesaplanır.	texture_mean texture_se texture _worst
3	Perimeter	Hücrelerin çevrelerinin ölçüm değerlerinin ortalama (mean), standart hata (se) ve en kötü değeri (worst) hesaplanır.	perimeter_mean perimeter_se perimeter_worst
4	Area	Hücrelerin yüzey alanlarının ortalama (mean), standart hata (se) ve en kötü durum değeri (worst) hesaplanır.	area_mean area_se area_worst
5	Smoothness	Komşu hücrelerin yarıçap uzunluklarının ortalaması (mean), standart hatası (se) ve en kötü değeri (worst) hesaplanır.	smoothness_mean smooothness_se smoothness_worst
6	Compactness	perimeter^2 / area – 1,0	compactness_mean compactness_se compactness_worst
7	Concavity	Hücrelerin etrafındaki girinti ve çıkıntıların ortalaması (mean), standart hatası (se) ve en kötü değeri (worst) hesaplanır.	concavity_mean concavity_se concavity_worst
8	Concave points	Hücrelerin etrafındaki girinti ve çıkıntı alanlarının sayısına göre ortalama (mean), standart hata (se) ve en kötü değer (worst) hesaplanır.	concave points_mean concave points_se concave points_worst
9	Symmetry	Hücrelerin elips şeklindeki değişimlerinin ortalaması (mean), standart hatası (se) ve en kötü değeri (worst) hesaplanır.	symmetry_mean symmetry_se symmetry_worst
10	Fractal dimension	coastline approximation - 1	fractal dimension_mean fractal dimension_se fractal dimension_worst

Veri kümesinde yer alan 30 özniteliğe ait örneklerin iyi huylu ya da kötü huylu etiket durumuna göre dağılım grafikleri Şekil 2'de sunulmuştur.

Veri kümesindeki 30 özniteliğin birbiriyle nasıl ilişkili olduğunu ve iki değişkenli ilişkileri görmek için korelasyon haritası Şekil 3'de gösterilmiştir. Şekilde görüldüğü gibi fractal_dimension_mean, texture_se gibi bazı özniteliklerin diagnosis (teşhis) alanı ile korelasyonları çok düşük değerlerdedir. Bu gibi özniteliklerin veri kümesinden çıkarılması veri kümesinin boyutunu azaltarak modellerin performansını artırmaktadır.

Şekil 3'e göre sırasıyla concave point_worst (0,79), perimeter_worst (0,78), radius_worst (0,78), concave points_mean (0,78), perimeter_mean (0,74), area_worst (0,73), area_mean (0,71) öznitelikleri diagnosis (teşhis) alanı ile en yüksek korelasyona sahip alanlar olarak görülmektedir. Kullanılan farklı öznitelik seçim yöntemlerine göre seçilen öznitelikler değişkenlik gösterebilmektedir.

Şekil 2. Veri kümesinde iyi ve kötü huylu dağılımları (Benign and malicious distributions in the dataset)



Şekil 3. Özniteliklere ait korelasyon değerleri (Correlation values of attributes)

2.2. Öznitelik Seçim Yöntemleri (Feature Selection Methods)

Yüksek boyutlu veriler makine öğrenimi için sorun teşkil edebilmektedir. Bu tür verilere dayanan tahmine dayalı modeller aşırı uyum riskini taşıyabilmektedir. Bu sebeple öznitelik seçim yöntemleri ile veri kümeleri üzerinde öznitelik sayısı azaltılabilmektedir. Ayrıca, özniteliklerden birçoğu gereksiz olabilir ve bu da tahmin doğruluğunun bozulmasına yol açabilir. Çalışmada kullanılan öznitelik seçim yöntemleri bu başlık altında kısaca açıklanmıştır [30]. Özellikle medikal alanda veri setleri üzerinde öznitelik seçimi yöntemleri tercih edilmektedir [31, 32].

2.2.1. Temel bileşenler analizi (Principal component analysis-PCA)

PCA, bir veri kümesinin özniteliklerini temel bileşenler (Principal Component-PC) adı verilen ilişkisiz öznitelikler kümesi olarak adlandırılan yeni bir duruma dönüştürmek için kullanılan istatistiksel bir yöntemdir. Bu sebeple PCA değişkenliğin çoğunu koruyarak, bir veri kümesinin boyutunu azaltmak amacıyla kullanılabilmektedir. Temel Bileşen Analizinin amacı, hedef değişkenler hakkında herhangi bir ön bilgi olmaksızın değişkenler arasındaki en önemli kalıpları veya ilişkileri korurken bir veri kümesinin boyutluluğunu azaltmaktır [30].

2.2.2. Tek değişkenli öznitelik seçimi (Univariate feature selection)

Tek değişkenli öznitelik seçiminde (Univariate feature selection), özniteliğin hedef değişkeniyle ilişkisinin gücünü belirlemek için her öznitelik tek tek incelenir. Bu yöntemde farklı özniteliklere göre sıralanmış bir öznitelik listesi döndürülür [33]. Bu algoritmada k kadar en iyi özniteliği seçen selectkBest yöntemi sıklıkla tercih edilmektedir.

2.2.3. Özyinelemeli öznitelik seçimi (Recursive feature elimination)

Özyinelemeli öznitelik seçimi (Recursive feature elimination-RFE), boyut azaltmak için kullanılan bir özyineleme yöntemdir. Bu yöntem mevcut öznitelikleri her iterasyonda bir veya daha fazla olacak şekilde azaltmayı amaçlamaktadır. Özyinelemeli öznitelik seçiminde geriye doğru eleme işleminde farklı tekniklerin kullanılması mümkündür. Bu teknikler arasında destek vektör makineleri, rastgele orman, doğrusal regresyon yer almaktadır [34]. Özniteliklerin belirlenmesinde hesaplayıcı (estimator) olarak karar ağaçları, lojistik regresyon gibi farklı yöntemler kullanılmaktadır.

2.2.4. Fark enflasyon faktörleri (Variance inflation factors-VIF)

Variance Inflation Factors (VIF), tüm modelin varyansının yalnızca söz konusu özniteliğe sahip bir modelin varyansına oranını doğrudan ölçer. Bir özniteliğin dâhil edilmesinin, modeldeki özniteliklerin katsayılarının genel varyansına ne kadar katkıda bulunduğu ölçülür. VIF'nin 1 olması, özniteliğin diğer özniteliklerinden herhangi biriyle korelasyonunun olmadığını gösterir [28]. VIF yapay zeka uygulamalarında temel olarak veri kümesindeki özniteliklerin azaltılması için kullanılan bir yöntemdir.

2.3. Makine Öğrenmesi Yöntemleri (Machine Learning Methods)

Verilerin sınıflandırılması için kullanılan çok sayıda makine öğrenme yöntemleri vardır. Bu yöntemlerin birbirine göre avantaj ve dezavantajları vardır.

2.3.1. K en yakın komşu algoritması (K Nearest Neighboard - KNN)

K-en yakın komşu (KNN) sınıflandırıcısı sık kullanılan klasik bir yöntemdir. K-En Yakın Komşular (KNN) algoritması, sınıflandırma ve regresyon görevleri için kullanılan popüler bir makine öğrenme tekniğidir. Benzer veri noktalarının benzer etiketlere veya değerlere sahip olma eğiliminde olduğu fikrine dayanır. Bu yöntemde bazı öznitelik vektörleri tarafından temsil edilen bilinmeyen bir örnek öznitelik uzayında bir nokta olarak sınıflandırmak için uzaklık hesaplanır. Bu uzaklık hesabında oklid, manhattan gibi farklı teknikler kullanılabilmektedir. Sınıflandırma işleminde bulunulacak örnek veri noktasının bulunduğu sınıfın ve en yakın komşunun, k değerine (benzerliğe) göre belirlendiği bir denetimli makine öğrenme yöntemi olarak ifade edilmektedir. Yeni bir veri geldiğinde mevcut verilerle uzaklıkları hesaplanır ve en küçük uzaklık değerine sahip sınıfa dâhil edilir [35].

2.3.2. Naive bayes - Gaussian

Naive Bayes sınıflandırma algoritması, adını Matematikçi Thomas Bayes'den alan bir sınıflandırma/ kategorilendirme algoritmasıdır. Naive Bayes sınıflandırması olasılık ilkelerine göre tanımlanmış bir dizi hesaplama ile sisteme sunulan verilerin sınıfını tespit etmeyi amaçlar. Naive Bayes sınıflandırmasında eğitim için sunulan verilerin mutlaka bir sınıfı/kategorisi bulunmalıdır. Eğitilmiş veriler üzerinde yapılan olasılık işlemleri ile sisteme sunulan yeni test verileri, daha önce elde edilmiş olasılık değerlerine göre öğretilir ve verilen test verisinin hangi kategoride olduğu tespit edilmeye çalışılır [36].

2.3.3. Karar ağaçları (Decision Trees)

Karar ağacı, sınıflandırma ve regresyon görevleri için parametrik olmayan denetimli bir öğrenme, verilerin sınıflandırılması ve tahmin edilmesi için kullanılan bir makine öğrenimi algoritmasıdır. Bu algoritma, bir ağacın dalları gibi bir dizi karar düğümüne dayanır ve her düğümde bir öznitelik test edilir. Test sonucuna göre, ağaç farklı dallara ayrılır ve sonunda bir tahmin veya sınıflandırma yapılır [37].

2.3.4. Destek vektör makineleri (Support vector machines –SVM)

Destek vektör makinesi algoritmasının amacı, N boyutlu bir uzayda (N - öznitelik sayısı) veri noktalarını belirgin bir şekilde sınıflandıran bir hiperdüzlem bulmaktır. İki veri noktası sınıfını ayırmak için en uygun hiperdüzlem bulunmaya çalışılır. Destek vektör makineleri (SVM'ler), sınıflandırma, regresyon ve aykırı değerlerin tespiti için kullanılan denetimli bir öğrenme yöntemidir [38].

2.3.5. Rastgele orman algoritması (Random forest algorithm)

Rastgele Orman algoritması denetimli bir sınıflandırma algoritmasıdır. Algoritmada birden fazla ağaçtan oluşabilmektedir. Rastgele Orman algoritması ile Karar Ağacı algoritması arasındaki temel fark, rastgele orman'da kök düğümü bulma ve düğümleri bölme işlemlerinin rastgele oluşuyor

olmasıdır. Bu algoritma ile sınıflandırma ve regresyon için, eğitim aşamasında çok sayıda karar ağacı oluşturarak probleme göre sınıf tahmini yapılmaktadır [39].

Topluluk öğrenme yöntemlerinden biri olan Rastgele Orman (RF) algoritmasında verilerden örnekler seçilerek rastgelelik sağlanır. Her ağacın aldığı kararlar birleştirilerek sonuçlandırılır. Rastgele orman verideki aşırı öğrenmeye ve eksik değerlere karşı güçlüdür [40]. Rastgele orman algoritmasında kullanılan önemli parametrelerden biri toplam kaç ağacın yer alacağıdır. Bu çalışmada kullanılan scikit-learn kütüphanesinde yer alan RandomForestClassifier fonksiyonunda n_estimators ağaç sayısını belirler [41].

2.4. Performans Metrikleri (Performance Metrics)

Çalışmada sınıflandırma alanında literatürde sıklıkla kullanılan doğruluk, duyarlılık, özgüllük, kesinlik ve f-skor performans parametreleri tercih edilmiştir. Bu metrikler kısaca açıklanmıştır.

- **Doğruluk:** Doğruluk değeri modelde doğru tahmin edilen alanların toplam veri kümesine oranı ile hesaplanmaktır (1) [42].

$$Doğruluk = (GP + GN)/(GP + GN + YP + YN) \quad (1)$$

- **Duyarlılık:** Duyarlılık ise pozitif olarak tahmin edilmesi beklenenlerin hangi sayıda pozitif olarak tahmin edildiğini göstermektedir (2) [43].

$$Duyarlılık = GP/GP + YN \quad (2)$$

- **Özgüllük:** Özgüllük, gerçek negatif oran olarak ifade edilir ve bir testte negatif sonuç veren gerçekten de negatif olan örneklerin oranıdır (3) [43].

$$Özgüllük = GN/GN + YP \quad (3)$$

- **Kesinlik:** Kesinlik positif olarak tahminlenen değerlerin gerçekten kaç tanesinin pozitif olduğunun göstergesidir (4) [43].

$$Kesinlik = GP/GP + YP \quad (4)$$

- **F-Skor**: F-puanı, bir sistemin kesinlik ve geri çağırma değerlerinin harmonik ortalamasıdır [43].

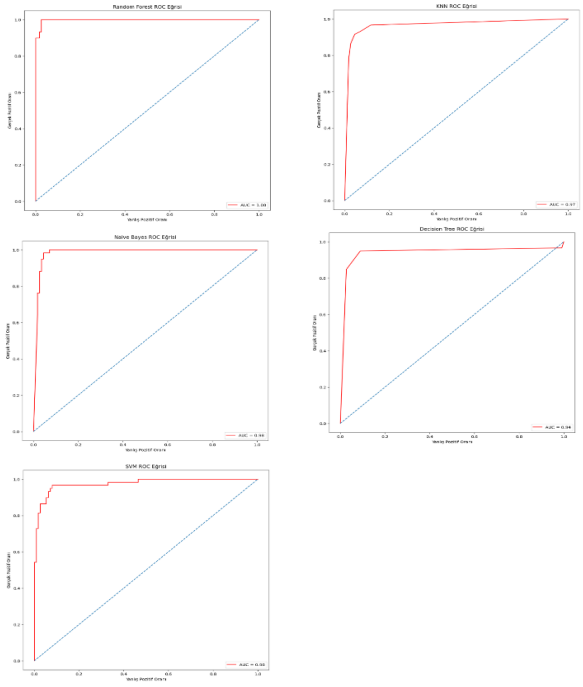
$$\text{F-Skor} = 2 * (Kesinlik * Duyarlılık)/(Kesinlik + Duyarlılık) \quad (5)$$

3. DENEYSEL SONUÇLAR VE TARTIŞMA (EXPERIMENTAL RESULTS AND DISCUSSION)

Bu çalışmada, göğüs kanserinin sınıflandırması için hem orijinal hem de öznitelik seçimi uygulanmış veri kümeleri kullanılmıştır. Tüm olası kombinasyonlar uygulanmaya çalışılarak en uygun öznitelik seçimi ve makine öğrenme yöntem ikilisi belirlenmeye çalışılmıştır. Veri kümeleri %70 eğitim %30 test olacak şekilde ayrılmışlardır.

3.1. Orijinal Veri Kümesi ile Elde Edilen Deneysel Sonuçlar (Experimental Results with Original Data Set)

Öncelikle orijinal veri kümesi öznitelik seçimine tabi tutulmadan farklı makine öğrenme algoritmaları ile eğitilmiş, ve test edilmiştir. Daha sonra her bir makine öğrenme modeline ait ROC eğrileri elde edilmiştir (Şekil 4). Her bir model için ağırlıklı ortalama kullanılarak elde edilen doğruluk (accuracy), duyarlılık, kesinlik ve özgüllük, f-skor (f-score) değerleri Tablo 2'de gösterilmiştir.



Şekil 4. Modellere ait ROC eğrileri (ROC curves of models)

Tablo 2. Orijinal veri kümesi ile elde edilen sonuçlar (Results with orginal dataset)

Yöntem	Parametre	Doğruluk %	Duyarlılık %	Özgüllük %	Kesinlik %	F-Skor %
KNN (K Nearest Neighbors)	K=5 manhattan	94,15	94,00	94,00	94,00	94,00
Naive Bayes	GaussianNB	95,91	96,00	96,00	96,00	96,00
Decision Tree	-	92,98	93,00	93,00	93,00	93,00
SVM	-	93,57	94,00	94,00	94,00	93,00
Random Forest	-	**97,66**	**98,00**	**98,00**	**98,00**	**98,00**

3.2. Öznitelik Seçimi ile Oluşan Veri Kümeleri ile Elde Edilen Deneysel Sonuçlar (Experimental Results Obtained with Data Sets Created by Feature Selection)

Orjinal veri kümesi üzerinde Principal Component Analysis (PCA), Variance inflation factors (VIF), Recursive feature elimination ve Univariate feature selection olmak üzere 4 farklı öznitelik seçme yöntemi uygulanarak veri kümesinin yeni versiyonları oluşturulmuştur. Daha sonra oluşturulan bu yeni veri kümeleri KNN, Naive Bayes, Decision Tree, SVM ve Random Forest makine öğrenme algoritmalarına uygulanarak en başarılı öznitelik seçim yöntemi ve makine öğrenme modeli tespit edilmiştir.

Tablo 3, 4, 5 ve 6 her bir öznitelik seçim yöntemine göre elde edilen sonuçları göstermektedir. Tablo 3, 4, 5 ve 6 ile sunulan sonuçlara bakıldığındaen yüksek doğruluk, duyarlılık, kesinlik ve f-skor değerleri Variance Inflation Factors öznitelik seçimi ve random forest yöntem ikilisi ile elde edilmiştir.

Variance Inflation Factors ile veri kümesinde yer alan 30 öznitelik 5'e düşürülmüştür. Bütün öznitelik indigeme yöntemlerinde öznitelik sayıları 5'e indirilerek karşılaştırma olanağı sağlanmıştır. Elde edilen sonuçlara göre doğruluk, duyarlılık, özgüllük, kesinlik ve f-Skor metriklerine göre en yüksek başarıyı gösteren Random Forest ve Variance Inflation Factors ikilisi karar destek sisteminin model bileşeni için seçilmiştir.

Variance Inflation Factors öznitelik seçimi yöntemine göre teşhis alanıyla en yüksek korelasyona sahip ilk 5 öznitelik 'radius_mean', 'perimeter_mean', 'radius_worst', 'perimeter_worst' ve 'area_mean' olarak belirlenmiştir. Bu alanlara göre veri setinin boyutu azaltılarak modelde kullanılmıştır.

Çalışmada Variance Inflation Factors ile elde edilen veri kümesinin Random Forest modeline uygulanması sonucu elde edilen karışıklık matrisine ait sonuçlar Tablo 7'de sunulmuştur. Şekil 5'te ise bu ikili ile elde edilen ROC eğrisi görülmektedir.

Tablo 3. PCA ile elde edilen veri kümesine ait deneysel sonuçları (Experimental results of the data set conducted with PCA)

Öznitelik Seçimi Yöntemi	Makine Öğrenme Yöntemi	Öznitelik Sayısı	Doğruluk %	Duyarlılık %	Özgüllük %	Kesinlik %	F-Skor %
PCA	KNN (k=5)	5	94,15	94,00	94,00	94,00	94,00
PCA	Naive Bayes	5	93,57	94,00	94,00	94,00	94,00
PCA	Decision Tree	5	96,49	96,00	97,00	97,00	97,00
PCA	SVM	5	93,57	94,00	94,00	94,00	94,00
PCA	Random Forest n_estimators=95	5	95,32	96,00	96,00	96,00	96,00

Tablo 4. Variance inflation factors ile elde edilen veri kümesine ait deneysel sonuçları (Experimental results of the data set conducted with variance inflation factors)

Öznitelik Seçimi Yöntemi	**Makine Öğrenme Yöntemi**	**Öznitelik Sayısı**	**Doğruluk %**	**Duyarlılık %**	**Özgüllük %**	**Kesinlik %**	**F-Skor %**
Variance Inflation Factors	KNN (k=5)	5	92,40	94,00	94,00	94,00	94,00
Variance Inflation Factors	Naive Bayes	5	94,74	95,00	95,00	95,00	95,00
Variance Inflation Factors	Decision Tree	5	96,49	96,00	96,00	96,00	96,00
Variance Inflation Factors	SVM	5	93,80	94,00	94,00	94,00	93,00
Variance Inflation Factors	**Random Forest**	**5**	**98,83**	99,00	99,00	99,00	99,00

Tablo 5. Univariate feature selection ile elde edilen veri kümesine ait deneysel sonuçları (Experimental results of the data set conducted with univariate feature selection)

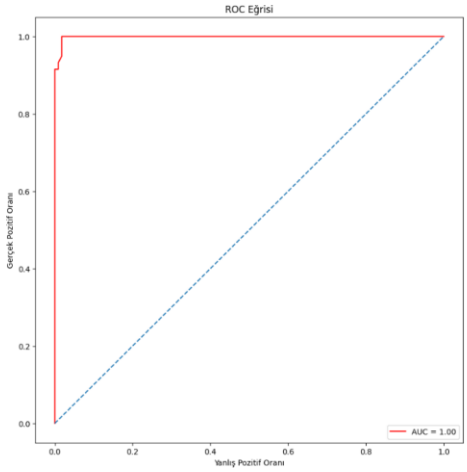
Öznitelik Seçimi Yöntemi	**Makine Öğrenme Yöntemi**	**Öznitelik Sayısı**	**Doğruluk %**	**Duyarlılık %**	**Özgüllük %**	**Kesinlik %**	**F-Skor %**
Univariate feature selection	KNN (k=5)	5	93,50	94,00	94,00	94,00	94,00
Univariate feature selection	Naive Bayes	5	91,81	92,00	92,00	92,00	92,00
Univariate feature selection	Decision Tree	5	92.10	92,00	92,00	92,00	92,00
Univariate feature selection	SVM	5	93,60	94,00	94,00	94,00	93,00
Univariate feature selection	Random Forest	5	95,32	95,00	95,00	95,00	95,00

Tablo 6. Recursive feature elimination (estimator: lojistik regresyon) ile elde edilen veri kümesine ait deneysel sonuçları (Experimental results of the data set conducted with recursive feature elimination)

Öznitelik Seçimi Yöntemi	**Makine Öğrenme Yöntemi**	**Öznitelik Sayısı**	**Doğruluk %**	**Duyarlılık %**	**Özgüllük %**	**Kesinlik %**	**F-Skor %**
Recursive feature elimination	KNN (k=5)	5	95,61	96,00	96,00	96,00	96,00
Recursive feature elimination	Naive Bayes	5	90,35	90,00	90,00	91,00	90,00
Recursive feature elimination	Decision Tree	5	94,74	95,00	95,00	95,00	95,00
Recursive feature elimination	SVM	5	92,11	92,00	92,00	92,00	92,00
Recursive feature elimination	Random Forest	5	94,91	95,00	95,00	95,00	95,00

Tablo 7. Variance inflation factors ve random forest ikilisi ile elde edilen deneysel sonuçlar (Experimental results obtained with variance inflation factors and random forest duo)

Doğruluk %	**Duyarlılık %**	**Özgüllük %**	**Kesinlik %**	**F Skor %**
98,83	96,00	99,00	99,00	99,00



Şekil 5. Recursive feature selection ve random forest ikilisi ile elde edilen ROC eğrisi (ROC curve obtained with recusive feature selection and random forest duo)

Tablo 8. Karşılaştırma sonuçları (Results of comparisons)

Referans	**Veri Seti**	**Öznitelik Seçimi Yöntemi – Öznitelik Sayısı**	**Makine Öğrenme Yöntemi**	**Doğruluk %**	**Hassasiyet %**	**Duyarlılık %**	**Kesinlik %**
Doğan vd. [10]	Breast Cancer Wisconsin Diagnostic	PCA	KNN (k=3)	92,40	94,50	-	93,64
Mohammed vd. [24]	Breast Cancer Wisconsin Diagnostic	-	SVM	97,70	-	**97,70**	**97,70**
Lavanya ve Rani [6]	Breast Cancer Wisconsin Diagnostic	Symmetric UncertAttributesetEval - 8	CART	94,72	-	-	-
Obaid vd., [25]	Breast Cancer Wisconsin Diagnostic	-	SVM	98,10	-	-	-
Agarap [26]	Breast Cancer Wisconsin Diagnostic	-	SVM	96.09	-	-	-
Rasol vd., [44]	Breast Cancer Wisconsin Diagnostic	RFS-15	SVM	98,68	-	98,22	98,95
Çalışma: Orijinal Veri Seti ile	Breast Cancer Wisconsin Diagnostic	-	Random Forest	97,66	98,00	98,00	98,00
Çalışma: VIF Öznitelik seçimi ile	**Breast Cancer Wisconsin Diagnostic**	**VIF - 5**	**Random Forest**	**98,83**	**99,00**	**99,00**	**99,00**

Literatürde göğüs kanserinin teşhisi ile ilgili çok sayıda çalışma bulunmaktadır. Bu çalışmalarda tomografi görüntüleri ya da sayısallaştırılmış öznitelik verileri kullanılmaktadır. Bu çalışmada sayısallaştırılmış öznitelik değerlerinden oluşan Breast Cancer Wisconsin (Diagnostic) Data Set tercih edildiği için literatürde bu veri kümesinin kullanıldığı çalışmalar üzerinde durulmuştur.

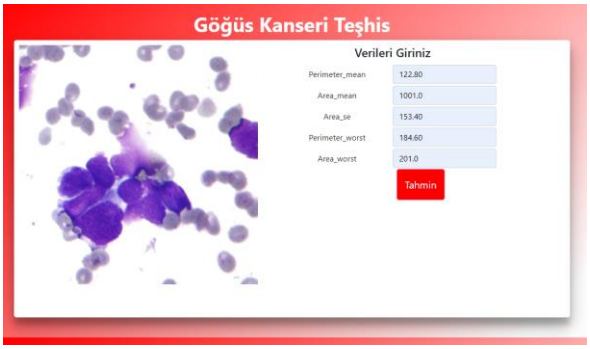
Tablo 8'de Breast Cancer Wisconsin (Diagnostic) Data Set isimli veri kümesini kullanan çalışmalara ait sonuçlar, bu çalışmada elde edilen sonuçlarla birlikte karşılaştırılmıştır.

4. GELİŞTİRİLEN KARAR DESTEK SİSTEMİ (DEVELOPED DECISION SUPPORT SYSTEM)

Karar destek sistemlerinin sağlık alanında kullanımı giderek yaygınlaşmaktadır. Karar destek sistemleri model, arayüz ve veri olmak üzere üç temel bileşene sahiptir. Çalışma kapsamında geliştirilen Karar destek sisteminin veri bileşenini kullanılan Breast Cancer Wisconsin (Diagnostic) Data Set adlı veri kümesi oluşturmaktadır.

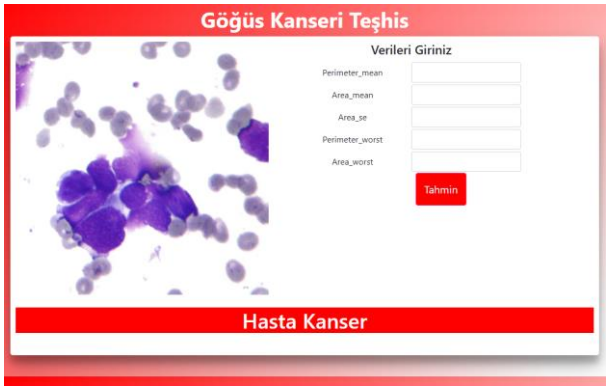
Öncelikle Python kodlama dilinde sade basit bir web ara yüzü oluşturmak için Flask kütüphanesi tercih edilmiştir. Flask, bir web uygulamasının tüm temel özelliklerini sunan bir mikro çerçevedir (framework). Python ile birlikte kullanılan Flask kütüphanesi ile Python uygulamaları için web arayüzleri oluşturulabilmektedir [45].

Model bileşeninde ise makine öğrenme yöntemleri yer almaktadır. Model bileşeninde makine öğrenmesinin yer almasıyla göğüs kanseri teşhisinde yüksek başarı sağlayan bir karar destek sistemi sunulmuştur. Şekil 6'da görüldüğü gibi ara yüz üzerinden 5 parametre girilerek tahmin sonucu elde edilebilmektedir. Bu 5 parametre, öznitelik seçimi sonucu teşhis (diagnosis) sütunuyla en yüksek korelasyona sahip ilk 5 özniteliktir. Bu değerler girildikten sonra "Tahmin" butonu kullanılarak model üzerinden elde edilen sonuç görülebilmektedir. Flask ile oluşturulan ara yüz render web sitesi kullanılarak deploy edilmiştir.



Şekil 6. Web ara yüz giriş sayfası (Web interface login page)

Web ara yüzünde tahmin butonuna basıldıktan sonra model ile elde edilen sonuç metin üzerinden kullanıcıya sunulmaktadır. Şekil 7'de görüldüğü gibi 5 parametreye göre elde edilen sonuç kullanıcıya tahmin butonunun altında sunulmaktadır.



Şekil 7. Web ara yüz sunumu (Web interface presentation)

5. SONUÇ VE ÖNERİLER (RESULTS AND SUGGESTIONS)

Çalışmada öznitelik seçimi ve makine öğrenme yöntemi ikilisinden en yüksek doğruluk (accuracy), duyarlılık (recall), özgüllük (sprecificity), kesinlik (precision) ve f-skor (f-score) değerlerinin elde edildiği Variance Inflation Factors öznitelik seçimi ve random forest makine öğrenme yöntemi web ara yüzünde tercih edilmişlerdir. Random forest algoritması bir düğümü bölerken en önemli özelliği aramak yerine, rastgele bir özellik alt kümesi arasından en iyi özelliği arar. Özniteliklerin sınıflandırılmasında literatürde en başarılı algoritmalar arasında yer almaktadır. Variance Inflation Factors ise birden fazla değişken arasındaki korelasyonu belirlemede etkili bir yöntemdir. Orijinal veri kümesinde random forest yöntemi ile ulaşılabilen en yüksek doğruluk değeri % 97,66 olurken, Variance Inflation Factors

öznitelik seçimi ve random forest yöntemiyle elde edilen doğruluk değeri %98,83 olarak gözlemlenmiştir. Variance Inflation Factors yöntemiyle öznitelik sayısı 30'dan 5'e düşürülmüştür. Bu sonuçlara göre öznitelik seçimi ile veri kümesinin boyutu azaltılmış ve doğruluk, duyarlılık, Özgüllük, kesinlik ve f-skor değerlerinde düşüş gözlemlenmemiştir. Öznitelik seçimi kullanılarak mevcut veri kümesinin boyutu azalmakta ve modelin daha hızlı çalışabilmesi sağlanmaktadır.

Çalışmanın web ortamına yüklenerek kullanıcıların erişimine açılabilmesi için Flask kütüphanesi ile bir web ara yüzü oluşturulmuştur. Flask ile web tabanlı bir ara yüz oluşturulmasının en büyük avantajı bu web sitesinin internet ortamına yüklenmesiyle erişilebilirliğinin artmasıdır. Bu çalışmada ücretsiz hosting hizmeti veren render.com adlı site kullanılmıştır. Sonuç olarak farklı öznitelik seçimi ve makine öğrenmesi kombinasyonları ile daha yüksek başarıyı daha az öznitelikle elde edecek çözümler sonraki çalışmalarda elde edilebilir. Bu çalışma farklı öznitelik seçimi ve makine öğrenme tekniklerinin karşılaştırması açısından önemli olabilecektir. Çalışmada veri kümesinde yer alan öznitelik sayısı 30'dan 5'e düşürülerek veri kümesinin boyutu düşürülmüştür. Yeni veri kümesi ile doğruluk (accuracy) % 1,17, kesinlik (precision) %1, duyarlılık (recall) %1, özgüllük (specificity) %1 ve f-skor (f-score) %1 artış göstermiştir.

Kanserin erken teşhisi hastaların tedavi edilebilmesinde önemli bir etkendir. Bu çalışma ile kanserin erken teşhisinde karar vericilere yardımcı olacak bir sistem sunulmaktadır. Kanser tedavisinde hastaları yoran kemoterapi gibi yöntemlerin yanı sıra Nano/mikromotor gibi yöntemler de geliştirilmektedir. Nano/mikromotorlar enerjiyi harekete dönüştürebilen küçük makineler olarak tanımlanabilir [46]. Sunulan çalışma teşhis başarısını artırması, tedavinin erken başlaması ve en uygun öznitelik seçimi ve yapay zeka yöntem ikilisini bulmaya çalışması ile birlikte literature katkı sağlayacaktır.

ETİK STANDARTLARIN BEYANI (DECLARATION OF ETHICAL STANDARDS)

Bu makalenin yazarı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

The authors of this article declare that the materials and methods used in their studies do not require ethics committee approval and/or legal-specific permission.

YAZARLARIN KATKILARI (AUTHORS' CONTRIBUTIONS)

Cihan AKYEL: Deneyleri yapmış, sonuçlarını analiz etmiş ve makalenin yazım işlemini gerçekleştirmiştir.

He conducted the experiments, analyzed the results and wrote the article.

Bünyamin CİYLAN: Veri setlerini düzenlemiş ve makalenin yazım işlemini gerçekleştirmiştir.

He organized the data sets and wrote the article.

Hüseyin POLAT: Çalışmada kullanılan tablolar, grafikler ve şekillerle ilgili düzenlemeleri yapmıştır. Çalışmada kullanılan metodların oluşturulmasında katkı sağlamıştır.

He made arrangements for the tables, graphs and figures used in the study. He contributed to the creation of the methods used in the study.

ÇIKAR ÇATIŞMASI (CONFLICT OF INTEREST)

Bu çalışmada herhangi bir çıkar çatışması yoktur.

There is no conflict of interest in this study.

KAYNAKLAR (REFERENCES)

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA A Cancer Journal for Clinicians. 2021; 2021: 209–249.

[2] Koçak HS, Olçar E, Güngörmüş Z. Birinci Derece Yakını Meme Kanserli Kadınların Korku Düzeyinin Erken Tanı Davranışlarına Etkisi. Hemşirelik Bilimi Dergisi. 2022; 6: 22-29.

[3] Altındağ Bayrak E, Kırcı P, Ensari T, Seven E, Dağtekin M. Göğüs Kanseri Verileri Üzerinde Makine Öğrenmesi Yöntemlerinin Uygulanması. Journal of Intelligent Systems: Theory and Applications. 2022; 5: 35-41.

[4] Pantel P. Breast cancer diagnosis and prognosis. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=78266df15a9754b7661f1f01722f9f4aea4244fb

[5] McMorran J, Crowther DC. Fine needle aspiration cytology (breast),

https://link.springer.com/book/10.1007/978-3-031-26900-4

[6] Lavanya D, Rani DKU. Analysis of feature selection with classification: Breast cancer datasets. Indian Journal of Computer Science and Engineering (IJCSE). 2011; 2: 756-763.

[7] Tamer HY. Akıllı Şehirlerde Veri Yönetimi Yaklaşımları. Abant Sosyal Bilimler Dergisi. 2022; 22: 519-534.

[8] Koçak A, Ergün PMA. Sağlıkta veri kalitesi ve veri madenciliği uygulamaları. Disiplinlerarası Yenilik Araştırmaları Dergisi. 2023; 3: 23-30.

[9] Demir, F. Ultrason RF Sinyallerinden Göğüs Kanserinin Derin Öğrenme Tabanlı Yaklaşımlarla Tespit Edilmesi. Fırat Üniversitesi Mühendislik Bilimleri Dergisi. 2022; 34: 761-768.

[10] Doğan, H, Tatar A, Tanyıldızı AK, Taşar B. Breast Cancer Diagnosis with Machine Learning Techniques. Bitlis Eren Üniversitesi Fen Bilimleri Dergisi. 2022; 11: 594-603.

[11] Bozkurt Keser S, Keskin K. Ağırlıklı Oy Tabanlı Topluluk Sınıflandırma Algoritması ile Göğüs Kanseri Teşhisi. Mühendislik Bilimleri ve Araştırmaları Dergisi. 2022; 4: 112-120.

[12] Erdem E, Aydin T. Göğüs Kanseri Histopatolojik Görüntü Sınıflandırması. Bilişim Teknolojileri Dergisi. 2022; 14: 87-94.

[13] Talo M. Meme Kanseri Histopatalojik Görüntülerinin Konvolüsyonal Sinir Ağları ile Sınıflandırılması. Fırat Üniversitesi Mühendislik Bilimleri Dergisi. 2019; 31: 391-398.

[14] Spanhol F, Oliveira E, Petitjean C, Heutte L. Breast cancer histopathological image classification using Convolutional Neural Networks. International Joint Conference on Neural Networks (IJCNN). 2016; 32: 2560-2567.

[15] Han Z, Wei B, Zheng Y, Yin Y, Li K, Li S. Breast cancer multi-classification from histopathological images with structured deep learning model. Scientific reports, 2017; 7: 4172-4182.

[16] Alom Z, Yakopcic C, Taha M, Asari K. Breast Cancer Classification from Histopathological Images with Inception Recurrent Residual Convolutional Neural Network. J Digit Imaging, 2019; 45: 1-13.

[17] Kahya, AAM, Al-Hayani W, Algamal ZY. Classification of breast cancer histopathology images based on adaptive sparse support vector machine. Journal of Applied Mathematics and Bioinformatics. 2017; 7: 1-15.

[18] Gupta V, Bhavsar A. Breast Cancer Histopathological Image Classification: Is Magnification Important. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW): Proceedings. 2017: 769-776.

[19] Dandıl E, Serin Z. Derin Sinir Ağları Kullanarak Histopatolojik Görüntülerde Meme Kanseri Tespiti. Avrupa Bilim ve Teknoloji Dergisi. 2020; Ejosat Özel Sayı: 451-463.

[20] Narin A, Kefeli SK. Meme Kanserinin Evrişimsel Sinir Ağı Modelleriyle Tespitinde Farklı Görüntü Büyütme Oranlarının Etkisi. Karaelmas Fen ve Mühendislik Dergisi. 2020; 10: 186-194.

[21] Akalın B, Veranyurt Ü. Sağlıkta Dijitalleşme Ve Yapay Zekâ. SDÜ Sağlık Yönetimi Dergisi. 2022; 2: 128-137.

[22] Hoşgör H, Güngördü H. Sağlıkta Yapay Zekanın Kullanım Alanları Üzerine Nitel Bir Araştırma. Avrupa Bilim ve Teknoloji Dergisi. 2022; 35: 395-407.

[23] Purkuloğlu E, Ün A, Yürürdurmaz F. Hemşire Karar Destek Sistemleri Uygulamaları. Hacettepe Sağlık İdaresi Dergisi. 2019; 22: 491-514.

[24] Mohammed TR, Al-Aaraj H, Rubbai YSY, Arabyat MM. Diagnosis of Breast Cancer Pathology on the Wisconsin Dataset with the Help of Data Mining Classification and Clustering Techniques. Applied Bionics and Biomechanics. 2022; 2022: 1-9.

[25] Obaid OI, Mohammed MA, Ghani MKA, Mostafa A, Taha F. Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. International Journal of Engineering & Technology. 2018; 7: 160-166.

[26] Agarap AFM. On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. The 2nd International Conference on Machine Learning and Soft Computing (ICMLSC '18): Proceedings. 2018: 1-5.

[27] Salem N, Hussein S. Data dimensional reduction and principal components analysis. Procedia Computer Science. 2019; 161: 292-299.

[28] Marcoulides KM, Raykov T. Evaluation of Variance Inflation Factors in Regression Models Using Latent Variable Modeling Methods. Educational and Psychological Measurement. 2019; 79: 874-882.

[29] Çetin Taş İ. An Applied Analysis of Breast Cancer Diagnosis By Using Different Methods. Abant Sağlık Bilimleri ve Teknolojileri Dergisi. 2022; 2: 72-87.

[30] Howley T, Madden MG, O'Connell M, Ryder AG. The Effect of Principal Component

Analysis on Machine Learning Accuracy with High Dimensional Spectral Data. Knowledge-Based Systems. 2006; 19: 209-222.

[31] Parlak, B, Uysal, AK. On feature weighting and selection for medical document classification. In Developments and advances in intelligent systems and applications. Springer International Publishing. 2018; 718: 269-282.

[32] Parlak, B, Uysal, AK. On classification of abstracts obtained from medical journals. Journal of Information Science. 2020; 46: 648-663.

[33] Subho RH, Chowdhury R, Chaki D, Islam S, Rahman M. A Univariate Feature Selection Approach for Finding Key Factors of Restaurant Business. IEEE Region 10 Symposium: Proceedings. 2019: 605-610.

[34] Niquini FGF, Branches AMB, Costa JFCL, Moreira GC, Schneider CL, Araújo FC, Capponi LN. Recursive Feature Elimination and Neural Networks Applied to the Forecast of Mass and Metallurgical Recoveries in A Brazilian Phosphate Mine. Minerals. 2023; 13: 748-759.

[35] Hu LY, Huang MW, Ke SW et al. The distance function effect on k-nearest neighbor classification for medical datasets. SpringerPlus. 2016; 5: 1-9.

[36] Anand MV, KiranBala B, Srividhya SR, Kavitha C, Younus M, Rahman H. Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer. Mobile Information Systems. 2022; 2022: 1-7.

[37] Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015; 27: 130-135.

[38] Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing. 2020; 408: 189-215.

[39] Breiman L. Random Forests. Machine Learning. 2001; 45: 5-32.

[40] Şahin H, İçen D. Application of Random Forest Algorithm for the Prediction of Online Food Delivery Service Delay. Turkish Journal of Forecasting. 2021; 5: 1-11.

[41] Saygılı A. Classification and Diagnostic Prediction of Breast Cancers via Different Classifiers. International Scientific and Vocational Studies Journal. 2018; 2: 48-56.

[42] Powers, D, Powers A. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies. 2011; 2: 2229-3981.

[43] Sokolova, M, Japkowicz, N, Szpakowicz, S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. AI 2006: Advances in Artificial Intelligence. Lecture Notes in Computer Science. 2006; 4304: 1015-1021.

[44] Rasool A, Bunterngchit C, Tiejian L, Islam MR, Qu Q, Jiang Q. Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis. International Journal of Environmental Research and Public Health. 2022; 19: 1-19.

[45] Aslam FA, Mohammed HN, Lokhande PS. Efficient Way Of Web Development Using Python And Flask, International Journal of Advanced Research in Computer Science. 2015; 6: 54-57.

[46] Türker A, Bülbül YE, Öksüz A, Yurdabak Karaca G. Kanser Teşhis ve Tedavisinde Nano/mikromotor Teknolojisi. Gazi University Journal of Science Part C: Design and Technology. 2023; 11: 652-672.