**BOZOK ÜNİVERSİTESİ**
**MÜHENDİSLİK MİMARLIK FAKÜLTESİ**
**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**

# AKCİĞER KANSERİNİN KARAR AĞAÇLARI ALGORİTMASI İLE TESPİT EDİLMESİ

**ÖRÜNTÜ TANIMAYA GİRİŞ**

**Faide KARATAŞ-16008118013**
**Özlem ÖZKAYA-16008118023**

**Dr. Öğr. Üyesi Muhammet Emin ŞAHİN**

**2021-2022 BAHAR DÖNEMİ**

**İçindekiler**

## Giriş

Akciğer kanseri, akciğerde oluşan kötü huylu tümörün dokulara veya organlara yayılarak vücuda zarar vermesi durumunda ortaya çıkmaktadır. En çok rastlanan kanser türleri arasındadır. Ölüme neden olan kanser türleri arasında ilk sıralarda yer almaktadır.

Akciğer kanserinin erken tespit edilmesi halinde tedavisi ve iyileşme olasılığı artmaktadır. Bu nedenle günümüzde kanserin erken teşhisi için yapılan çalışmalar önem kazanmaktadır.

Bu çalışmada, csv formatında ve bilgisayarlı tomografilerinden oluşan iki farklı veri seti kullanılmıştır. Lung Cancer veri setinde hastalardan alınan, kanser genetiğini etkileyen 13 faktör ve teşhis bilgisi bulunmaktadır. Irak-Onkoloji Eğitim Hastanesi/Ulusal Kanser Hastalıkları Merkezi (IQ-OTH/NCCD) akciğer kanseri veri setinde ise kötü huylu ve normal tümörlerin bilgisayarlı tomografi görüntüleri bulunmaktadır. Her iki veri setine de makine öğrenmesi algoritmalarından Karar Ağaçları algoritması uygulanarak Lung Cancer veri setinde hangi hastanın akciğer kanseri olabileceği hangisinin olamayacağı tahmin edilmiş, IQ-OTH/NCCD akciğer kanseri veri setinde ise tümör türleri tespit edilmiştir.

Lung Cancer veri setinde sayısal ve kategorik veriler mevcuttur. Makine öğrenmesi yöntemini uygulamak için kategorik verilerin sayısallaştırılması gerekmektedir. Bu nedenle veri seti LabelEncoder yöntemi ile sayısallaştırılmıştır. Ardından korelasyon grafiği gösterilmiştir. Karar Ağaçları algoritmasını uygulamadan önce standardizasyon işlemi yapılmıştır. Standardizasyon işlemi ortalama değerin 0, standart sapmanın ise 1 değerini aldığı, dağılımın normale yaklaştığı bir metoddur.

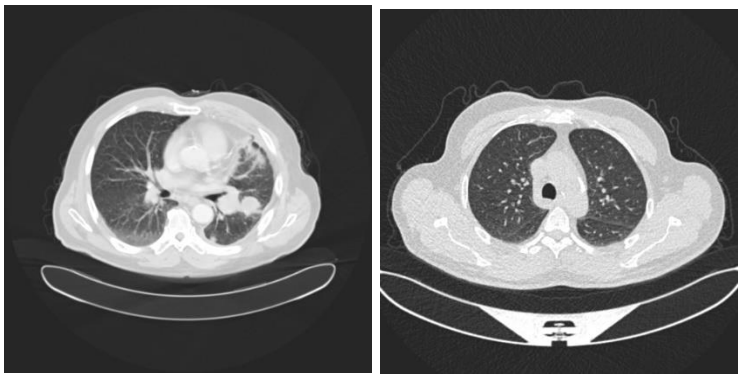
## Materyal ve Yöntem

### Veri Seti

Projede erişime açık olan IQ-OTH/NCCD- Lung Cancer Dataset ve LC.csv veri setleri kullanılmıştır [1] [2]. İlk veri seti için kanserli (lung) ve normal akciğer röntgeni görüntülerini içermektedir. 416 normal ve 561 lung olmak üzere toplam veri sayısı 977 adettir. Şekil 1’de ilk veri setine ait Şekil 2’de diğer veri setine ait görseller bulunmaktadır.

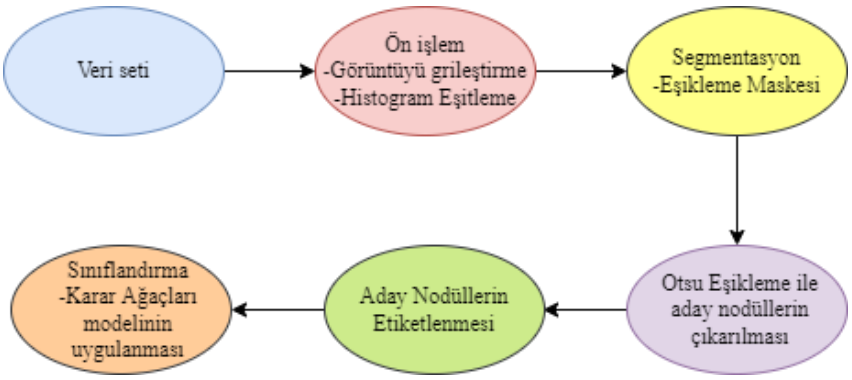
| | id | diagnosis | TP53 | PTEN | EGFR | KRAS | FGFR1 | ALK | METex14 | MET | BRAF | PIK3CA | ROS1 | HER2 | RET |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1001 | B | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **1** | 1002 | B | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **2** | 1003 | B | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **3** | 1004 | B | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 1005 | M | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **5** | 1006 | B | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **6** | 1007 | B | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

Şekil 1: Veri setinden örnek görüntüler(Lung Cancer Dataset)



Şekil 2: Veri setinden örnek görüntüler(LC.csv)

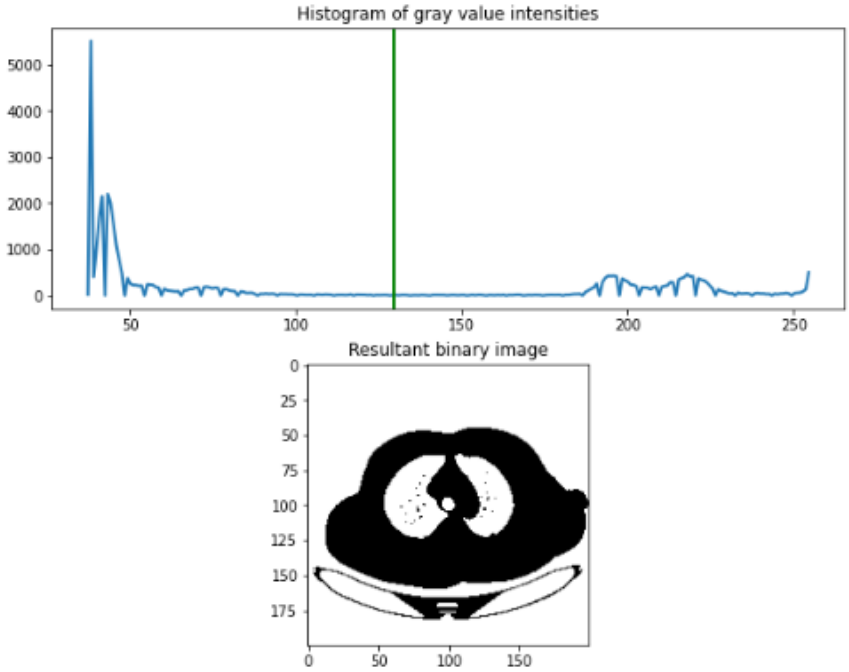
Projede CT görüntüleri üzerinde tümör dokularının tespitini yapmak için 5 ana bölümden oluşan işlemler görüntüler üzerinde uygulanmıştır. Yapılan tüm işlemlerin akış diyagramı Şekil 3'de verilmiştir.




Şekil 3: Tümör Tespit Uygulaması Akış Diyagramı

**Ön İşlem**

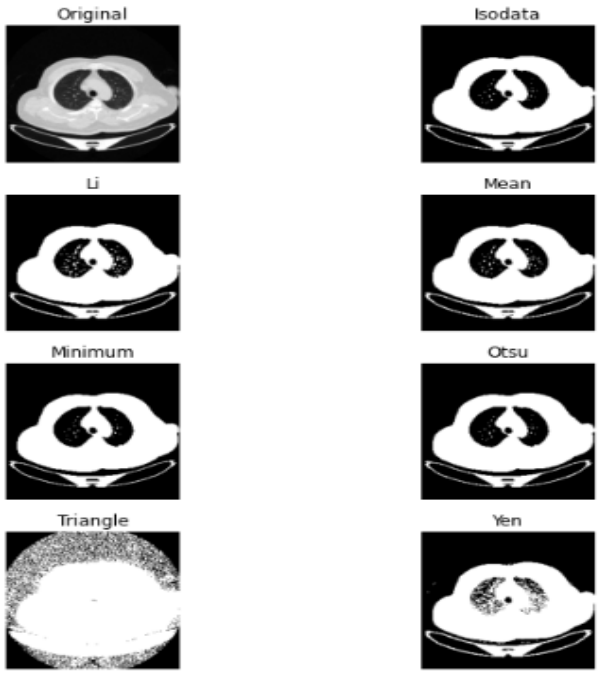
Jpg formatında okutulan veri seti öncelikle segmentasyon işlemleri ve gri formata çevirme işlemleri gerçekleştirilmiştir. Çünkü double türünde veriler ile çalışılmaktadır. İlk olarak histogram eşitleme işlemi ve boyut eşitleme işlemleri gerçekleştirilmiştir. Şekil 4'de elde edilen görsel ve histogram grafiği mevcuttur.




Şekil 4: Ön işleme gerçekleşmiş veri ve histogramı

**Akciğer Segmantasyonu**

Eşikleme, gri tonlamalı bir görüntüden ikili bir görüntü oluşturmak için kullanılır burada kullanılmıştır. “skimage.filters” , kütüphanesi ile sağlanan eşikleme algoritmalarını değerlendirmek için bir fonksiyon içerir ve birden fazla yöntemi aynı anda değerledirler. Şekil 5’de gösterilmiştir. En güzel soncu “otsu” yönteminde olmuştur.



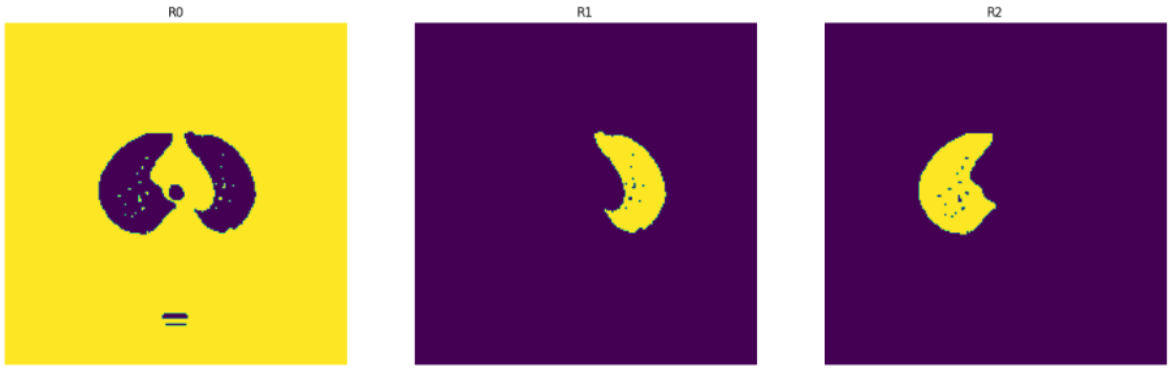

Şekil 5: skimage.filters kullanımı

**Özellik Çıkarımı**

Öznitelik çıkarma adımında sınıflandırma işlemi için gerekli öznitelikler elde edilmiştir. Sınıflandırma ve makine öğrenimi araçları doğrudan görüntüler üzerinde değil, görüntü işleme uygulamalarındaki görüntülerde bulunan özellikler üzerinde çalışır. Özellik çıkarma işlemleri için özellikleri çıkarılacak parçalar segmente edilmiş akciğer görüntüleri üzerinden otsu eşikleme metodu ile seçilmiştir.

Elde edilen bu görüntü üzerindeki her bir parça görüntü etiketleme yöntemi ile otomatik olarak etiketlendirilmiştir. Etiketleme işlemi sonrasında her görüntü içindeki her bir parça için ayrı ayrı sınırlayıcı kutu haritaları çıkartılmıştır. Bu sınırlayıcı kutu haritaları lojik formatta olup tümör ve nodüllerin görüntülerinin çıkarılması için birer maske olarak kullanılmıştır.

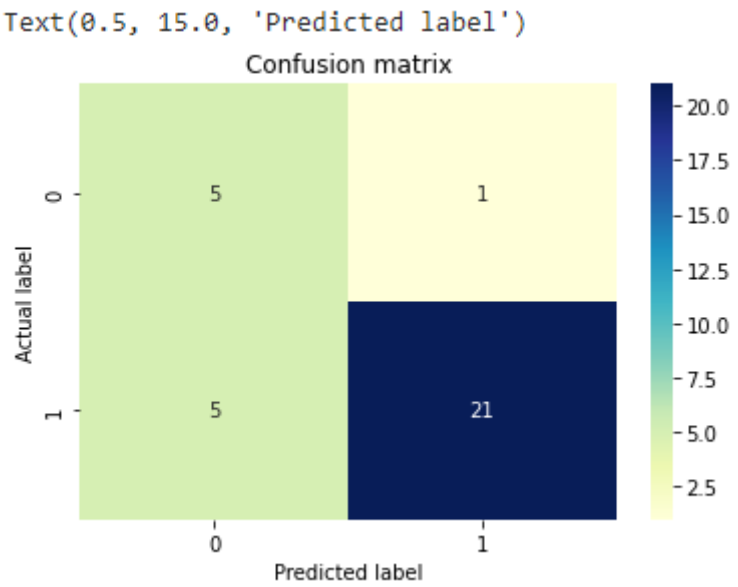
Ortaya çıkan bu görüntünün her bir parçası, görüntü etiketleme yöntemi kullanılarak otomatik olarak etiketlenir. Etiketleme işleminden sonra her görüntünün her parçası için ayrı bir sınırlayıcı kutu oluşturulur. Bu sınırlayıcı kutu ile tümör görüntülerini çıkarmak için maskeler çıkarılır. Maskelenmiş görüntüleri Şekil 6'da gösterilmektedir.



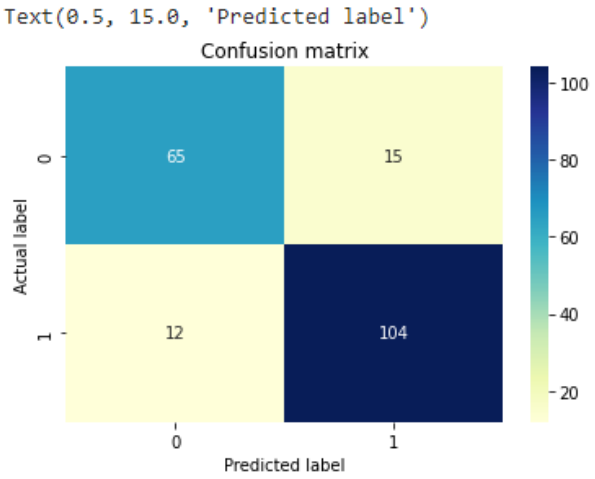

Şekil 6: Maskelenmiş ve etkilenmiş görüntü

**Sınıflandırma**

Çalışmanın son adımında Karar Ağaçları algoritması uygulanıp eğitim LC.cvs veri seti doğruluğu 0.927, test doğruluğu 0.812 ve Lung Cancer Dataset doğruluğu 0.90, test doğruluğu 0,86 olarak elde edilip sonuçlar matrisleri Şekil 7-8'de gösterilmiştir.




Şekil 7: LC.cvs  Confusion matrix




Şekil 8: Lung Cancer Dataset  Confusion matrix

**Karar Ağaçları**

Karar ağaçları, büyük boyuttaki veri setlerini probleme uygun şekilde karar kuralları uygulayarak daha küçük parçalara bölmek için kullanılan bir algoritmadır.

Karar ağaçları düğümler, dallar ve yapraklar olmak üzere üç temel yapıdan oluşmaktadır. Ağaç yapısında, her özellik bir düğüm ile temsil edilir. Dallar ve yapraklar, ağaç yapısının diğer bileşenleridir. KA yapısının temel çalışma prensibi, veri setini küçük parçalara bölerek en kısa sürede elde etmektir. KA'da, ağaçtaki her bir düğüm bir sınıfı temsil eder veya düğümdeki test verilerini oluşturan çıktılara göre örnek bölümler ayrılarak bir test bölümü oluşturur. Parçalanan her alt küme, bir alt ağaçla çözülerek yeni bir alt sınıflandırma problemi ortaya çıkaracaktır ve çalışma yapısı böyle devam edecektir. Yaprak adı verilen düğümler, sonuç düğümünün sınıfını

içerir. Yaprak düğümü olmayan noktalar ise karar düğümleri olarak adlandırılır. Bu karar düğümleri yeni bir özellik oluşturur ve bu özelliklerin olabilecek her değeri için aynı ağacın dallarını parçalayarak başka bir karar ağacı oluşturur.

Başlangıçta mantık, istatistik ve yönetim amacıyla türetilen karar ağaçları, günümüzde metin madenciliği, bilgi çıkarma, örüntü tanıma ve makine öğrenimi gibi birçok alanda oldukça etkili bir şekilde kullanılmaya başlanmıştır.

**Gini**

Gini katsayısı veya Gini indeksi, İtalyan istatistikçi Corrado Gini tarafından 1912'de geliştirilen bir ölçüdür. Katsayı değerleri, 0 ile 1 aralığındadır. 0, eşitliği temsil eder ve 1 ise eşitsizliği temsil eder. Gini katsayısı, rastgele seçilen bir özelliğin hangi sıklıkta yanlış tespit edildiğini ölçmek için kullanılan bir indekstir. Problemlerde düşük Gini indeksi olan bir özellik seçilmelidir. Gini katsayısı kategorik hedef değişkeni için başarılı veya başarısız olarak çalışır (0 veya 1). Gini indeks, yalnızca ikili (binary) bölmeleri gerçekleştirir. Yüksek gini indeksi homojenliği artırır. CART (Sınıflama ve Regresyon Ağacı) ikili bölmeler oluşturmak için gini yöntemini kullanır. Şekil 9'da formülü verilmiştir.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Şekil 9: Gini index formülü

**Entropi**

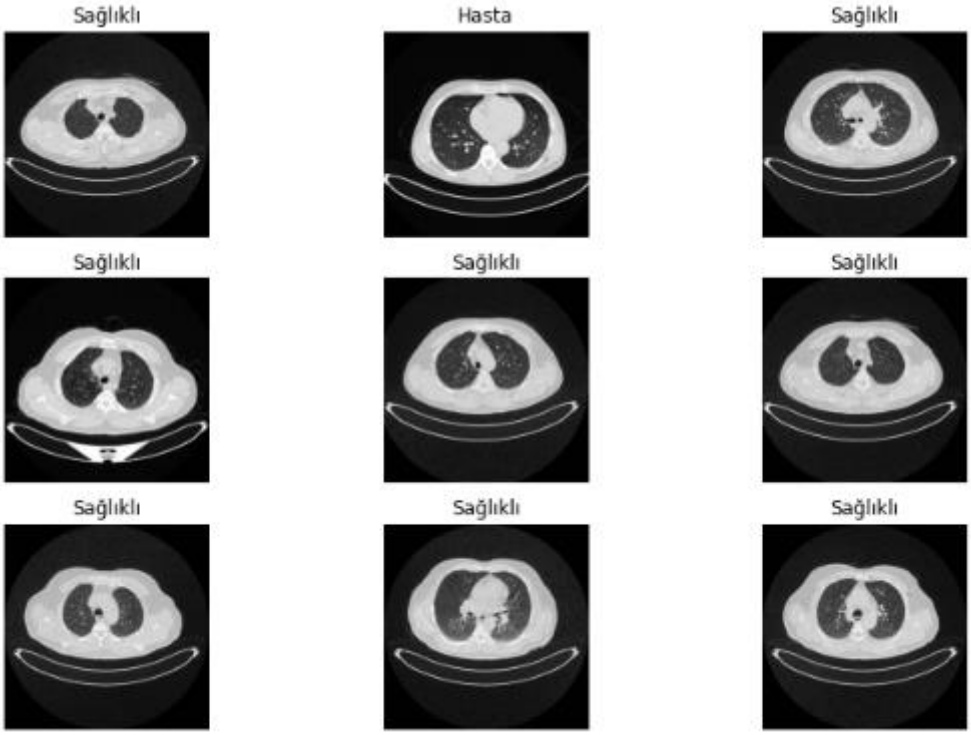
Entropi, rasgele bir özelliğin belirsizliğinin ölçüsüdür. Örneklerin bir koleksiyonunun saf olup olmadığını karakterize eder. Entropi ne kadar fazla olursa elde edilen bilgi de o kadar fazla olur. Şekil 10'da formülü verilmiştir.

$$Entropy(D_1) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

Şekil 10: Entropi formülü

## Sonuç

Elde edilen sınıflandırma sonuçlarına göre kanser dokularının tespiti yüksek başarı oranlarıyla test edilmiştir. Şekil 11' de gösterilmiştir.




Şekil 11: Kanser tespiti

**Kaynakça**

[2] “IQ-OTH/NCCD - Lung Cancer Dataset | Kaggle.” https://www.kaggle.com/datasets/adityamahimkar/iqothnccd-lung-cancer-dataset (accessed May 19, 2022).

[1] “Farazkhan0516/Lung-Cancer-Prediction-using-Machine-Learning: Project Material for predicting lung cancer in a patient with the help of machine learning.” https://github.com/Farazkhan0516/Lung-Cancer-Prediction-using-Machine-Learning (accessed May 20, 2022).