



**RADYOGENOMİK VERİ SETLERİ İÇİN YAPAY ZEKÂ
TABANLI BİR HASSAS TIP KARAR DESTEK SİSTEMİNİN
GELİŞTİRİLMESİ**

Abdulahap PINAR

BİYOİSTATİSTİK VE TIP BİLİŞİMİ ANABİLİM DALI

**Tez Danışmanı
Doç. Dr. Emek GÜLDOĞAN**

Doktora Tezi – 2025

**T.C.
İNÖNÜ ÜNİVERSİTESİ SAĞLIK
BİLİMLERİ ENSTİTÜSÜ**

**RADYOGENOMİK VERİ SETLERİ İÇİN YAPAY ZEKÂ TABANLI BİR
HASSAS TIP KARAR DESTEK SİSTEMİNİN GELİŞTİRİLMESİ**

Abdolvahap PINAR

Biyoistatistik ve Tıp Bilişimi Anabilim Dalı

Doktora Tezi

Tez Danışmanı

Doç. Dr. Emek GÜLDOĞAN

Tez Jüri Üyeleri

Prof. Dr. Cemil ÇOLAK

Prof. Dr. Hakan PARLAKPINAR

Doç. Dr. Emek GÜLDOĞAN

Prof. Dr. Memiş BOLACALI Doç.

Dr. Mehmet Onur KAYA

Bu araştırma YÖK 100/2000 Doktora Bursu kapsamında Biyoistatistik ve Tıp Bilişim Anabilim Dalı Biyoinformatik öncelikli alanda desteklemektedir.

MALATYA

2025

T.C.
İNÖNÜ ÜNİVERSİTESİ
Sağlık Bilimleri Enstitüsü Müdürlüğüne

ETİK BEYANI

İnönü Üniversitesi Sağlık Bilimleri Enstitüsü Tez Yazık Kurallarına uygun olarak Doç. Dr. Emek GÜLDOĞAN danışmanlığında hazırlayıp sunduğum Radyogenomik Veri Setleri İçin Yapay Zekâ Tabanlı Bir Hassas Tıp Karar Destek Sisteminin Geliştirilmesi Doktora tezim için elde ettiğim verileri, bilgileri, belgeleri akademik ve etik kurallar çerçevesinde elde ettiğimi, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu, tezimde yararlandığım eserlere bilimsel kurallara uygun atıfta bulunarak kaynak gösterdiğimi, tezimin özgün olduğunu, tezimin çalışma ve yazımında patent ve telif haklarını ihlal edici bir davranışımın olmadığını, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabul ettiğimi beyan ederim.

09/05/2025

Abdulahap PINAR

İÇİNDEKİLER

ÖZET	viii
ABSTRACT	ix
SİMGELER VE KISALTMALAR DİZİNİ.....	x
ŞEKİLLER DİZİNİ	xi
TABLolar DİZİNİ.....	xii
1. GİRİŞ.....	1
2. GENEL BİLGİLER	5
2.1. Akciğer Kanseri.....	5
2.2. Genomik.....	6
2.2.1. Yapısal Genomik.....	7
2.2.2. Fonksiyonel Genomik.....	7
2.3. Radyogenomik	8
2.4. Görüntü İşleme	9
2.5. Yapay Zekâ	10
2.6. Makine Öğrenmesi.....	11
3. MATERYAL VE METOT	13
3.1. Veri Seti.....	13
3.2. Klinik Verilerde Ön İşleme Yaklaşımları	13
3.2.1. Veri Temizleme	13
3.2.2. Veri Dönüştürme.....	14
3.2.3. Veri Entegrasyonu.....	14
3.2.4. Değişken Seçme Yöntemi.....	15
3.2.5. Boyut İndirgeme	15
3.2.6. Veri Bölme	15
3.3. Görüntü Verilerinde Ön İşleme Yaklaşımları.....	16

3.4. Evrişimsel Sinir Ağları	18
3.4.1. Random Forest (RF)	19
3.4.2. Aşırı Gradyan Artırma Algoritması	20
3.4.3. Categorical Boosting Algoritması (CatBoost)	21
3.4.4. Destek Vektör Makineleri	22
3.5. Biyoistatistiksel Analizler	22
3.6. Geliştirilen Web Tabanlı Yazılım	22
3.6.1. Ana Sayfa Sekmesi	23
3.6.2. DICOM Görüntülerinin Okunması	23
3.6.3. Görüntü Analizi ve Sınıflandırma Sekmesi	24
4. BULGULAR	26
5. TARTIŞMA	39
6. SONUÇ VE ÖNERİLER	43
KAYNAKLAR	44
EKLER	52
EK-1. Özgeçmiş	52
EK-2. Etik Kurul Onayı	53

TEŞEKKÜR

Akademik eğitimim ve çalışmalarım boyunca bilgi ve deneyimleri ile bana destek olan değerli danışman hocam Sayın Doç. Dr. Emek GÜLDOĞAN'a, eğitimim boyunca her daim sabırla desteğini esirgemeyen çalışmalarımda ve önerileriyle bana ışık tutan Sayın Dr. Öğr. Üyesi Ahmet Kadir ARSLAN ve Sayın Prof. Dr. Cemil ÇOLAK'a sonsuz teşekkür ve saygılarımı sunarım. Doktora eğitimim süresince her daim yanımda olan, Sayın Prof. Dr. Harika Gözde GÖZÜKARA BAĞ, Arş. Gör. Fatma Hilal YAĞIN, Arş. Gör. Şeyma YAŞAR, Arş. Gör. Zeynep KÜÇÜKAKÇALI ve Arş. Gör. İpek BALIKÇI ÇİÇEK'e teşekkür ve saygılarımı sunarım. Ayrıca, hayatımda derin izler bırakan ve 6 Şubat 2023 depreminde aramızdan ayrılan sevgili arkadaşım Pınar YILMAZ'ı minnetle anıyor, bana kattığı her şey için teşekkür ederim. Her türlü desteğini esirgemeyen canım aileme teşekkür ederim

ABDULVAHAP PINAR

ÖZET

Radyogenomik Veri Setleri İçin Yapay Zekâ Tabanlı Bir Hassas Tıp Karar Destek Sisteminin Geliştirilmesi

Amaç: Bu çalışmanın amacı, KHDAK hastaların biyomedikal görüntüleri ve ilgili klinik veriler kullanılarak derin öğrenme yöntemine dayalı DICOM görüntülerden süper piksel segmentasyon, otsu eşikleme ve hastalığın referans konumu belirlenmiştir. Görüntülerden alınan nicel veriler klinik veriler ile entegre edildi. Biyobelirteçlerin belirlenmesi için çeşitli makine öğrenme yöntemleri uygulanarak sınıflandırma performanslarını, klinik veriler, klinik veriler+görüntü verileri ve modele ilişkin iyileştirme oranlarını belirlemektir.

Materyal ve Metot: Çalışmada kullanılan klinik veri setinde, EGFR mutasyonuna sahip 43, mutasyon olmayan 168 ve KRAS mutasyonuna sahip 38, mutasyon olmayan 173 birey olmak üzere her bir mutasyonda 211'er KHDAK bulunmaktadır. İşlenen görüntü sayısı 2.231 ve her bir görüntü VGG16 derin öğrenme modeli ile 25.088 değişken çıkarıldı. Sınıflandırma algoritmalarından, XGBoost, CatBoost, Random Forest, DVM modelleri uygulanmıştır.

Bulgular: Küçük hücreli Dışı Akciğer kanseri hastaların EGFR ve KRAS mutasyon durumları biyomedikal klinik verilerde anlamlı farklılıklar bulunmuştur. Random forest yöntemi ile en önemli 50 değişken seçilerek model eğitiminde kullanılmıştır. En iyi sınıflandırma model performansının XGBoost, CatBoost yöntemlerine ait olduğu belirlenmiş ve bu modellere ilişkin sınıflandırma performansları doğruluk, dengelenmiş doğruluk, kesinlik, duyarlılık, F1-skoru, ROC-AUC metrikleri ile değerlendirilmiş ve bunlar sırasıyla, (0.965 ± 0.015) , (0.954 ± 0.021) , (0.953 ± 0.024) , (0.994 ± 0.007) , (0.973 ± 0.011) ve (0.990 ± 0.005) olarak hesaplanmıştır.

Sonuç: Gerçekleştirilen çalışmada makine öğrenmesi yöntemleri sonuçlarından hareketle, görüntü verilerinde KRAS mutasyon durumunu tahmin etmede etkili olduğu XGBoost ve CatBoost modelleri daha başarılı performans gösterdiği saptanmıştır. EGFR mutasyon durumunu tahmin etmede CatBoost daha başarılı performans göstermiştir.

Anahtar Kelimeler: Akciğer Kanseri, Biyomedikal, Dijital Görüntüleme, Makine Öğrenme

ABSTRACT

Development of an Artificial Intelligence Based Precision Medicine Decision Support System for Radiogenomics Data Sets

Aim: This study's objective was to use biomedical images of NSCLC patients and associated clinical data to use a deep learning approach to identify super pixel segmentation, herbaceous thresholding, and disease reference location from DICOM images. Clinical data and quantitative information from image were combined. Biomarker identification, classification performance, clinical data, clinical data + image data, and model improvement rates were all assessed using a variety of machine learning techniques.

Materials and Methods: The study's clinical dataset included 211 NSCLCs with each mutation, including 38 persons with a KRAS mutation and 173 without, and 43 individuals with an EGFR mutation and 168 without. A total of 2,231 photos were analyzed, and 25,088 features were retrieved from each image using the VGG16 deep learning model. XGBoost, CatBoost, Random Forest, and SVM models were utilized among the classification techniques.

Results: EGFR and KRAS mutation status of lung cancer patients were found to have significant differences in biomedical clinical data. The most important 50 features were selected by random forest method and used in model training. It was determined that the best classification model performance belonged to XGBoost, CatBoost methods and the classification performances of these models were evaluated with accuracy, balanced accuracy, precision, sensitivity, F1-score, ROC-AUC metrics, respectively, (0.965 $\pm$ 0.015), (0.954 $\pm$ 0.021), (0.953 $\pm$ 0.024), (0.994 $\pm$ 0.007), (0.973 $\pm$ 0.011) and (0.990 $\pm$ 0.005), respectively.

Conclusion: The results of the study's machine learning techniques indicate that both XGBoost and CatBoost models can accurately identify the presence of KRAS mutations in image data, with XGBoost and CatBoost models performing better than the others. CatBoost was better at predicting the presence of EGFR mutations.

Keywords: Biomedical, Digital Imaging, Lung Cancer, Machine Learning.

SİMGELER VE KISALTMALAR DİZİNİ

AUC	: ROC Eğrisi Altında Kalan Alan
BT	: Bilgisayarlı Tomografi
CNN	: Evrişimsel Sinir Ağı
DÖ	: Derin Öğrenme
DVM	: Destek Vektör Makinesi
EGFR	: Epidermal Büyüme Faktörü Reseptörü
KHDAK	: Küçük Hücreli Dışı Akciğer Kanseri
KRAS	: Kirsten Sıçan Sarkomu Viral Onkogen Homoloğu
ML	: Makine Öğrenmesi
PET	: Pozitron Emisyon Tomografisi
RF	: Random Forest
YSA	: Yapay Sinir Ağları
YZ	: Yapay Zeka

ŞEKİLLER DİZİNİ

S ek il No	Sayfa No
Şekil 1.1. DÖ yöntemi ile radyomik, genomik ve radyogenomik arasındaki ilişki	3
Şekil 3.1. BT görüntülerinde süper piksel segmentasyonu ve sonuçları	17
Şekil 3.2. Otsu eşikleme sonrası görüntünün maskelenmesi ve referans noktalarının belirlenmesi	18
Şekil 3.3. Random Forest algoritması karar ağaç yapısı	20
Şekil 3. 4. Ana sayfa sekmesi	23
Şekil 3. 5. DICOM görüntülerinden meta verilerin çıkarılması	24
Şekil 3. 6. Modellere ilişkin çıktılar	25
Şekil 4.1. KHDAK hastalarına ilişkin DICOM görüntülerine ilişkin segmentasyon, maskeleme ve kontur analizi	30
Şekil 4.2. EGFR mutasyonuna ilişkin en önemli değişkenlerin belirlenmesi	31
Şekil 4.3. ROC eğrilerini kullanarak makine öğrenmesi model performanslarının karşılaştırılması	33
Şekil 4.4. Akciğer görüntülerinde segmentasyon, maskeleme ve kontur analizi	34
Şekil 4.5. KRAS mutasyonuna ilişkin en önemli değişkenlerin belirlenmesi	35
Şekil 4.6. ROC eğrilerini kullanarak makine öğrenimi model performanslarının karşılaştırılması	36

TABLÖLAR DİZİNİ

Tablo No	Sayfa No
Tablo 4.1. Kategorik verilere ilişkin tanımlayıcı istatistikler	26
Tablo 4.2. Nicel verilere ilişkin tanımlayıcı istatistikler	29
Tablo 4.3. EGFR mutasyon durumuna göre sınıflandırma performansına ilişkin metrikler	32
Tablo 4.4. KRAS mutasyon durumuna göre sınıflandırma performansına ilişkin metrikler	35
Tablo 4.5. EGFR ve KRAS mutasyonlarının sınıflandırılmasına ilişkin tablo	37

1. GİRİŞ

Kanser arařtırmaları, 1990'ların bařında ortaya ıkan genomik devriminden bu yana hassas tıp tedavilerini mmkn kılmak amacıyla hastalıkların nedenlerini genetik boyutta arařtırmaya odaklanmıřtır. İnsan genom projesi tamamlandıktan sonra gen ve proteinlerin tasarımı, gen bilimi ve gen rollerin iřlevsel dzeylerine evrilmiřtir. Kanser Genom Atlası (TCGA) projesi gibi birok kanser projesi, genom trnde transkriptom, epigenomik ve proteomik veriler elde etmek iin alıřmalar yapıldı (1, 2). Bylelikle genetik analiz yntemlerinin bařlıca riskleri ve ortaya ıkabilecek olası komplike belirtiler tařıyan iřlemler geliřtirilmiřtir. Genetik analizi, tmr deęiřkenine ait gruplar ii ve gruplar arası heterojen olmasından kaynaklı olarak doku biyopsi rnekleri genetik deęiřimleri doęru bir řekilde ifade edemeyebilir (3). Vakalar oęu zaman heterojen yapılı lezyonların ok az bir blmnden elde edildięinden ve lezyonun anatomik iřlevi, fizyolojik etkenlerini doęru olarak temsil etmeyebilir (4-6). Tedaviye alınan bireylerin tedavi boyunca elde edilen ıktılar iin birden fazla dokuyu incelemek mmkn olmaz (7). Klinik uygulamalarda ve tıp bilimlerinde grntleme olduka nemlidir. Tıbbi grntlemenin iřlevi, kiřselleřtirilmiř hassas tıp yn ile hastalığın teřhisinde baskın bir ara olarak hızlı bir řekilde geliřmektedir (8, 9). Hassas tıp ve byk veriler ile yapılan alıřmaların artmasıyla beraber alanında birok uzman ‘‘Radyogenomik’’ kavramının nemini ne srmektedir. Radyogenomik, tıbbi grntleme ve genomik veriler arasındaki benzer iliřkileri geliřtirmek ve radyomik ile genomuęın eksik ynlerini kombine ederek ok lekli iliřkilerin geliřtirilmesine olanak saęlamaktadır (10). Radyogenomik ile fenotip veya genotip olarak bilinen gamet eřitlerinin tahminine ynelik yaygın kullanılan bir stratejidir. Sınıflandırma modelinin kullanılması iin radyogenomik yapının iki veya daha fazla grupta kmelere ayırmak ve sınıflandırma modelleri kullanılarak uygun kestirim modeli elde edilir. Radyogenomik sınıflandırma problemi iin model tahmini genel olarak gerek deęerlidir. Tahmin edilen performans deęerlerinin belirlenmesi iin doęru pozitif, doęru negatif, yanlış pozitif ve yanlış negatif deęerleri gz nne alınır (11, 12).

Dnya apında akcięer kanserine baęlı lmler ok yaygın bir řekilde grlmektedir (13). Akcięer tanısı tıbbi grntleme radyografi veya bilgisayarlı tomografi (BT) ile belirlenir ve genel olarak grntleme ile elde edilen ıktılar akcięerde

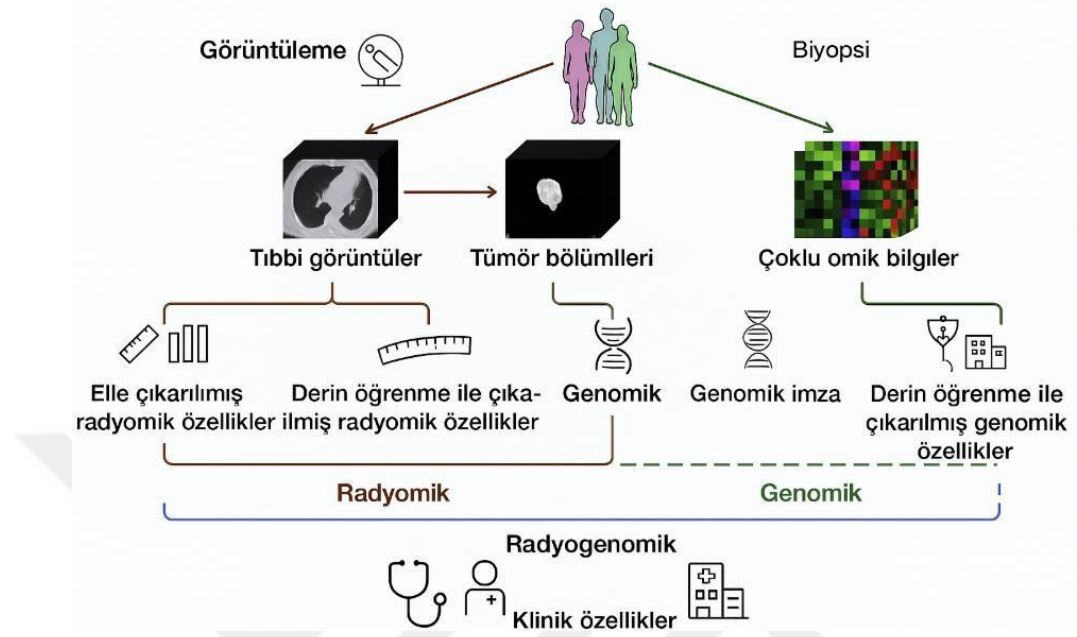
bulunan bir lezyonun varlığı ve bu lezyonu çevreleyen dokular ile (plevra, kaburgalar, hilus vb.) aralarındaki ilişki belirlenir. Belirlenen lezyonlar genel olarak kanser teşhisini belirlemek ve tümörün histolojik belirtileri küçük hücreli akciğer kanseri (KHAK), küçük hücreli dışı akciğer kanseri (KHDAK), gibi için biyopsi yapılır (13).

Akciğer kanserinin gen ifade profilleri ve radyogenomik bir haritanın oluşturulması amacıyla, 113 küçük hücre dışı kanser hastası BT teknolojisi ile görüntüler sunulmuş bir çalışmada, akciğer kanseri fiziksel ve biyokimyasal süreçlerde meydana gelen bozukluklarda Epidermal Büyüme Faktörü Reseptörü (EGFR) önemli derecede baskın olduğu çeşitli moleküller düzeylerini kapsayan pek çok sayıda metageni ortaya çıkmasında (Ribonükleik asit) RNA dizileme yöntemi kullanılarak gerçekleştirilmiştir. Bu nitelikler göz önüne alınarak metagenler arasındaki ilişkileri korelasyon yöntemiyle detaylı radyogenomik haritalar oluşturulmuştur (14). Radyogenomik bir çalışmada uygulanması gerekli olan temel adımlar şöyle sıralanır;

- ✓ Adım 1. Hastalıklı olan bölge bölümlere ayrılır ve ilgili hastalık üç boyutlu hale getirilir.
- ✓ Adım 2. Hastalığın hacminde verimli bilgi elde etmek için radyomik özellikler çıkarılır.
- ✓ Adım 3. Radyogenomik özellikleri geliştirmek için klinik, radyomik ve genomik nitelikler birleştirilir.
- ✓ Adım 4. Radyomik özellikleri bulunabilmesi için radyomik fenotipler ve genotipler arasında korelasyon katsayıları incelenir.
- ✓ Adım 5. Yaşam analizini tahmin etmek için radyogenomik modeller kullanılır. Şeklinde sıralanabilir (15).

Derin Öğrenme (DÖ) modelleri tıbbi görüntüleme analizi olarak kullanılan önemli bir yöntemdir. Sağlık bilimlerinin pek çok alanında özellikle tıbbi görüntü algılama ve görüntü bölütleme alanlarında derin öğrenme ağ mimarileriyle elde edilen modellere çokça ihtiyaç duyulmaktadır. Erken teşhis ve tedavi gibi alanlarda sıklıkla yararlanılır (16). DÖ'nün medikal hizmetlerde, tıbbi görüntüleme diğer gerçek dünya problemleriyle ilişkilendirildiğinde derin öğrenme ile belirtilen problemler daha yavaş ilerlemektedir. DÖ mimarilerin gelişmesinde ve tıbbi görüntüleme çalışmaları doğrultusunda büyümeyi etkileyen faktörler analiz edilerek BT teknolojisi uygulama

alanları araştırılmış ve ilgili alanlar vurgulanmıştır. DÖ ile radyomik, genomik ve radyogenomik arasındaki ilişki gösterilmiştir (Şekil 1.1) (17).



Şekil 1.1. DÖ yöntemi ile radyomik, genomik ve radyogenomik arasındaki ilişki

DÖ tekniği ile akciğer kanserinin radyogenomiği tespit etmenin en temel yolu, BT görüntülerinden elde edilen genomik biyobelirteçlerin değişim durumunu kestirmektir. BT cihazları ile mutasyon durumunun bir kestirimi ve optimal süreç için en iyi ağaç tabanlı yöntemin tespiti için epidermal (hücre bölünmesi, hücre büyümesi, hücre yaşaması, vb.) büyüme faktörü modellenerek evrimsel sinir ağı (CNN) modeli eğitilmiştir (18). Görüntü analizi akciğer düğümü ve kitle segmentasyonu uygulamalarında nitelik belirlemenin öğrenilmesi için tıbbi alanda CNN'ler kullanılarak büyük başarılar göstermiştir (19). Hassas tıp tedavilerine olanak sağlamak amacıyla hastalıkların nedenlerini genetik boyutta araştırmaya odaklanmıştır. Kanser hastaları dünya çapında ölüme neden olan hastalıklardan biri olarak açıklanmaktadır. Ölüme neden olan bu hastalığın diğer ölümcül hastalar gibi BT aracılığı ile görüntüleme analizi gerçekleştirilir. Kanser hastalığı verileri bu tür çalışmalar için destek görmekte ve veri setleri kamuya açık bir şekilde sunulmaktadır (20). Kamuya açık biyomedikal görüntülere ilişkin verileri karşılaştırmak için geliştirilen veri tabanları, önceki modelleri değerlendirerek yeni modeller üretmek için kullanıcıya kaynaklar sağlayarak görüntü analizi algoritmalarının geliştirilmesine katkıda bulunsa bile, bazı veri kümelerinin birden

fazla yerde dağıtılmış olması veya farklı terminolojiler kullanılarak indekslenmesi, modellerin güvenilir bir şekilde karşılaştırılması ve yeniden üretilebilirliği zorlaştırmaktadır. Akciğer Görüntü Veri tabanı Konsorsiyumu görüntü koleksiyonu sayesinde biyomedikal veri kümelerinin modellere ilişkin temel doğrulukların karşılaştırılması mümkündür (21, 22).

Bu tez çalışmasında, KHDAK için klinik ve radyogenomik görüntü veri setlerini çeşitli ön işleme süreçlerinden geçirerek işleyen ve sonrasında akciğer kanserine yönelik makine öğrenmesi (ML) tabanlı tahmin modelleri oluşturan, bir tıbbi karar destek sisteminin geliştirilmesi amaçlanmıştır.



2. GENEL BİLGİLER

2.1. Akciğer Kanseri

Akciğer kanseri, günümüz insan sağlığını tetikleyen en büyük hastalıklardan biridir. Kanserın erken tespit edilmesi durumunda tedavi yöntemi çok etkili olabilmektedir. Akciğer hastalığına bağılı olarak çeşitli hastalıkların ortaya çıkmasını engellen en önemli etken erken teşhis ve tedavinin etkin olmasıdır. Akciğer kanseri, yapılan bilimsel çalışmalara göre erkeklerde prostat kanserinden sonra kadınlarda ise meme kanserinden sonra sıklıkla görülen hastalıkların kansere bağılı ölümlerin başında gelmektedir (23). Akciğer kanseri genel kanser hastalıklarının arasında %13.4'ünü meydana getirirken, ölümlle sonuçlanan kanser hastalıklarının %28.4'ünde akciğer kanseri olduğu tespit edilmiştir. Amerika Birleşik Devletlerinde yapılan araştırmalara göre, akciğer kanseri hariç diğer kanser hastalıklarının beş yıllık sağkalım oranı %52 olup akciğer kanserinde ise bu oran %14 olarak tespit edilmiştir. Gelişen tedavi imkânlarına rağmen ölüm oranları artmıştır (24). Sigara içen bireylerde akciğer olma riski %10-20 oranında risk tespit edilmiş olup, bu durum akciğer kanseri olma riski geçmişte ailesinde akciğer kanserinin olabilme ihtimalinin yatkın olduğu gösterilmektedir (25, 26). Akciğer kanseri BT, PET ve göğüs radyografisi gibi görüntüleme yöntemleri ile akciğerde nodülleri tespit etmek için kullanılabilir. Bu teknikler arasında BT görüntüleme yöntemi PET ve göğüs radyografisi tekniklerine göre daha hassas olduğu için daha fazla tercih edilmektedir (27).

Nodüllerin tespiti, alanında uzman hekimler için zor bir durum olduğundan, analizi yapılacak olan BT taramaların sayısı azımsanmayacak kadar fazla olduğu için radyologlar açısından iş yükü önemli ölçüde artmaktadır. Hastalığın erken tespit edilmesi ve alınan önlemlerde artışın sağlanması için bilim adamları; uzman hekimlerin iş yükünü hafifletmek, hastalığa neden olan etkenin nicelik olarak azaltarak ilgili hastalığın teşhis edilmesinde netliği artırmak ve daha hızlı sonuç analizlerinin yanı sıra maliyetin azalması yönünde bilgisayarlı çözüm odaklı çalışmalar sürdürülmektedir (28, 29).

BT tarama yöntemi ile erken tanı, akciğer kanseri hastalarda ölüm oranlarını azaltabilir fakat bu yöntem maliyet açısından oldukça yüksek olması, gelir seviyeleri düşük ve orta düzeyde olan ülkeler açısından büyük risk taşımakta olup uzmanlar BT görüntüleme yöntemleri olmadan, sağlıklı vaka analizinin gerçekleştirilmemesi, yanlış

teşhis ve tedavi ile sonuçlanabilmektedir (29). Bu gibi durumların üstesinden gelebilmek için BT görüntü tarama yöntemlerinde lezyon saptama ve akciğer hastalığı sınıflandırma tekniklerini geliştirilmek için, son yıllarda BT görüntü tarama, erken teşhis ve tedavi yöntemi kapsamında hem uzman hekimlerin yükünü hafifletme hem de maliyet açısından avantajlı olmasından dolayı bilgisayar tabanlı daha hızlı ve sonuç odaklı çözümler üreten YZ yöntemleri, görüntü işleme, biyomedikal görüntü analizi yöntemleri araştırma alanı haline almıştır (30).

Akciğer kanserinin tespit edilmesinde kullanılan ve görüntü işleme için uygulanan adımlar sırasıyla; görüntü yakalama, görüntü iyileştirme, görüntü bölütleme ve değişken seçme adımları izlenmektedir. İlk aşamada normal ya da anormal görüntüleri BT taramaları ile elde edilir. İkinci adımda, en iyi nitelikteki taramayı elde etmek için görüntü geliştirme yöntemleri uygulanır. Üçüncü adımda, görüntü ayırma algoritması uygulanır ve son olarak görüntülerin normal veya anormal göstergeler içeren bölümlenmiş görüntüden analizi yapılacak değişkenler elde edilir (31).

2.2. Genomik

Genomik, organizmada bulunan genetik bilginin analiz edilmesi için biyolojik verinin sıralanması, haritasının çıkarılması ve bu materyaller ile DNA'nın yapısını ve işlevini inceleyen yenilikçi bir disiplindir. Büyük veri dosyalarının kaydedilmesi ve verinin işlenmesi için uygun veri tabanlarının oluşturulması, genomik verinin işlenmesi, araştırmacının istenilen veriye erişimini sağlayabilmesi ve güncel verilerin kaydedilmesi literatüre katkı sağlayan önemli bir adımdır (32). Genomik tıbbın ilerleyişine katkı sağlayan en büyük faktör sağlık bilimlerinde uzman hekimlerin genetik ve tıbbi büyük verilere erişebilmesine dayanmaktadır. Genomik araştırmaların büyük ölçüde başarılı sonuçlar elde edilmesinde hasta bireylerin ve toplumun duyarlılığı neticesinde genomik disiplinin başarısı ve sürekli ilerleyişi büyük önem taşımaktadır. Genomik bilginin bir araya getirilmesinde ve araştırmacılara verilerin paylaşılması konularında sorumlu kurum ve kuruluşlara duyulan güven neticesinde, genomik veriler ile günümüz sağlık alanında YZ ve derin öğrenme modellerinin gelişmesiyle birlikte, genomik verinin görselleştirilmesi ve analizinin yapılması, araştırmacının daha etkin sonuçlar elde edilmesini hızlandırmıştır (33). Genomik kavramını özetle tanımlayacak olursak:

- Hastalık taşıyan organizmayı tanımlar ve fenotip süreci ile ilişkili gen ifadelerini açıklar.

- Yapısal ve fonksiyonel genomik yapılarından yola çıkarak genler arasındaki ilişkiler hakkında araştırmacıya bilgi sunar.

Çeşitli organizmalara sahip genetik yapıları gen tabanında karşılaştırılmasına imkân sağlar (34).

Genomik çalışmalar yapısal ve fonksiyonel genomik olmak üzere iki ayrılır.

2.2.1. Yapısal Genomik

Yapısal genomik çalışmalar, genetik materyalin organizasyonunu, kromozomların yapılarını, genlerdeki değişimleri ve genetik bilginin yapısını, genetik hastalıkları belirlemek ve genetik tabanlı tedavi yöntemleri gibi çeşitli alanlarda bilgi vermeyi amaçlar (35). Yapısal genomik çalışmalarının ana temelleri açıklanmıştır (35):

Kromozom anomalileri, Yapısı açısından normal bir kromozomun yapısında meydana gelen değişikliği ifade eden genomik çalışmalardır. Bu genomik anomali değişim/değişimler genetik hastalığın meydana gelmesindeki etken olabilir. Yapısal genomik kromozomlardaki bu tür anomali durumları açıklamaya yardımcı olmaktadır.

Genomik translokasyonlar, Birbirinden ayrılan kromozom parçacıklarını farklı kromozomlara taşımasını sağlayan genomik yapıyı tanımlar. Bu yapısal farklılıklar, genlerin klasik işleyişini etkileyerek hastalıkların meydana gelmesini tetikler.

Kromozom haritalama, Yapısal genom çalışmaları, kromozomlardaki genetik bilgiyi açıklar ve elde edilen bilgi neticesinde genom haritasını oluşturmayı ifade eder. Kromozomların genetik bilgiyi taşımaları, gen haritalarının oluşturulması, dizi analizlerinin yapılması gibi yaklaşımlar yapısal genomik çalışmaların alt kümelerini oluşturan genetik bilimin önemli yaklaşımlarıdır. Bu yaklaşımlar herhangi bir organizmanın genetik yapısı hakkında bilginin verilmesine öncülük ederler (35).

2.2.2. Fonksiyonel Genomik

Fonksiyonel Genomik, genetik bilginin işlenmesi, düzenlenmesi ve bilginin analizini gerçekleştirmek amacıyla yapılan çalışmaları ifade eder. Genlerdeki değişimin fenotiplere olan etkisini anlamlandırmayı amaçlayan genomik çalışmalardır. Fonksiyonel genomik çalışmalarının ana temelleri aşağıda açıklanmıştır (36).

Gen ifade analizi, Bir organizmanın içerisindeki genlerinin ne kadar ifade edildiğini anlamak için yapılan çalışmaların bütününe verilen addır. RNA düzeyindeki ifade (transkriptom) ya da protein düzeyindeki ifade (proteom) analizi sağlar.

Fenotip ve genetik varyasyonlar arasındaki ilişkilerin incelenmesi için genetik değişimin fenotipi hangi düzeyde etkilediğini ifade etmek amacıyla yapılan çalışmalar olup ilgili hastalığın genetik boyutunu ve gen faktörünü belirler ve genetik belirteçlerin fenotip açısından etkilerini inceler.

Fonksiyonel genomik, genetik bilginin ifade edilmesindeki işlevselliğini anlamlandırmak ve genetik hastalığın anlaşılmasına, tedavi yönteminin belirlenmesinde ve geliştirilmesi için önemli bir faktör olup, bu çalışmalar genel olarak biyoinformatik, moleküler biyoloji, istatistik ve bilgisayar bilimleri gibi pek çok disiplinin birleştirir (36).

2.3. Radyogenomik

YZ sağlık alanında kullanılan “Radyomik” ve “Genomik”in entegre edilmesi ile meydana gelen radyogenomik, farklı bilim alanlarında kullanılmasının yanı sıra hassas tıp ve onkoloji alanlarında kullanılan son teknoloji bilimi olarak literatürde yerini almıştır. Radyogenomik, hassas tıp genomik veriler ile büyük hacimli radyolojik görüntülerin birleştirilmesi ile riskin sınıflandırılmasıdır. Klinik birimlerde hasta bireyler üzerinde yapılan birçok YZ çalışmaları yardımı ile modeller üretilir. Kanseri çalışmalarında, yaşam analizi tahminlerinde ve onkoloji alanında yapılan çalışmalarda YZ kapsamında radyogenomik sonuçlar büyük başarılar ortaya çıkarmıştır (37). Genotip veya fenotip değişkenlerin kestirim değerlerini elde etmek için radyogenomik yaygın kullanılan bir yöntemdir. Hedef değişkeni en az iki grup olarak kümelemek ve sınıflandırma modelini oluşturarak tahminler yapılmaktadır (11, 12). Radyomik ve radyogenomik işleyiş süreci için radyografik görüntüler işlenerek bu görüntülerden veriler oluşturulur. Birinci adımda, görüntü alma ve yeniden düzenlemek, ikinci adımda, görüntüdeki ilgili alanı bölümlendirmek, üçüncü adımda, faktör seçme ve son adımda ise tahmin sonuçlarını elde etmek için ML tekniklerine dayalı modelleme işlemlerini gerçekleştirmektir (38). Böylelikle YZ sağlık alanının çeşitli uygulamalarında başarılı hizmetler sunmaktadır. YZ tabanlı hem ML hem de derin öğrenme yöntemleri ile radyogenomik çalışmalarda model performans hassasiyeti daha illeri düzeye taşınır. Model performansını değerlendirirken parametre değeri ne kadar yüksek olursa YZ yöntemi o kadar iyi sonuç üretmektedir. Yapılmış radyogenomik çalışmaların büyük bir kısmında doğruluk ve ROC eğrisi altında kalan alan (AUC) göz önüne alınarak model yorumlanarak, duyarlılık ve özgüllük değerleri gibi metrikler kullanılır (37). YZ yöntemlerinin yanı sıra uzmanlar tarafından uygulanan gelişmiş tarama yöntemleri, tanı ve tedavide minimal invaziv prosedürlerin

kullanımı, stereotaktik ablatif gibi radyasyon uygulamaları, yeni hedefe yönelik ilaçlar ve immünoterapi dahil olmak üzere akciğer kanseri tedavisinde önemli atılımlar gerçekleştirilmiştir (39). Bu yeni tedavi seçeneklerinin mevcudiyeti, özellikle küçük hücreli dışı akciğer kanseri olanlarda artan sağkalım oranlarıyla ilişkilendirilmiştir. KHDAK için iki yıllık göreceli sağkalım oranı 2009-2010'da %34 iken 2015-2016'da %42'ye yükselmiştir (40). Hedefe yönelik tedavi veya immünoterapilere hak kazanan KHDAK hastaları, biyobelirteçlere bağlı olarak %15 ila %62,5 arasında değişen 5 yıllık sağkalım oranlarıyla daha uzun yaşamaktadır. Sonuç olarak, biyobelirteç testi, özellikle ilerlemiş veya metastatik hastalığı olan KHDAK hastalarında tedavi seçimine rehberlik etmek ve en iyi sonuçları elde etmek için gereklidir (40, 41).

2.4. Görüntü İşleme

Sağlık alanında tanı ve tedavi yönteminde sıklıkla kullanılan BT yöntemi, aynı zamanda enformatik, matematik, fizik gibi alanlarda olduğu gibi mühendislik alanlarında ve disiplinler arası yapılan araştırmalarda da uygulama alanları mevcuttur. Görüntü işleme; görüntülenecek cihaz aracılığı ile ölçümlerden elde edilen görüntü verisini, çalışmanın amacı doğrultusunda değişken seçme, görüntülerden anlamlı bilgiler elde etmek gibi adımlarının uygulanması sürecidir (42). Akciğer kanserinin teşhisi için sıklıkla kullanılan bir görüntüleme tekniği olan BT, lezyonların boyutu, şekli, yoğunluğu ve diğer özellikleri üzerinde analiz yapmak için kullanılır (43). Görüntü işleme yöntemleri, akciğer dokusu ve lezyonları bölütlemek amacıyla kullanılır. Bu işlem, lezyonların sınırlarını belirleyerek ve özelliklerini analiz ederek önemli bilgiler elde etmeyi amaçlar. Araştırmacılar otomatik lezyon tespiti performansını inceleyerek bu yöntemin etkinliğini vurgulamaktadır (44). Lezyonların şekli ve yoğunluğu, ML algoritmalarıyla sınıflandırmak için kullanılır. Bu yöntem, kanserli ve kanser öncesi lezyonları ayırtmak amacıyla desteklenmiştir (45). BT taramalarında otomatik lezyon tespitinin sonuçlarını değerlendirmek için görüntü işleme yöntemlerinin performans ölçülerinden hassalık, özgüllük ve diğer metrikler kullanılmıştır (46). Manyetik Rezonans (MR), genel olarak tıpta kullanılan görüntüleme tekniğidir. Canlı organizmanın içyapısını görüntülemek amacı ile kullanılmaktadır. Görüntüler, dokulardaki hidrojen atomlarının yoğunluğu ve hareketine bağlı olarak oluşturulur. Bu yöntemde radyasyon kullanılmaz; bunun yerine vücuttaki hidrojen atomlarının çekirdeklerine ait protonlar manyetik alana uyarılır. Canlı organizmalara herhangi bir zararı kanıtlanmamış olup, gebeler de bu gruba ilave

edilmiştir. Ancak organ gelişiminin gerçekleştiği ilk üç ayda MR çekimi önerilmez. Metal etkileşimi, mıknatıs veya metal protez taşıma, kalp pili kullanma, göz içinde yabancı cisim bulundurma, ateşli silah yaralanması geçirme veya kalıcı dövme durumları MR cihazına girmek konusunda riskli taşımaktadır (47).

2.5. Yapay Zekâ

YZ algoritması, Makinelerin verileri karşılaştırma yeteneği, sağlık sektöründe hasta takip planlaması, veriyi öğrenme, iletişim kurma, algılama, görüntü işleme, nesneleri oynatma ve yer değiştirme yeteneklerine sahiptir. Ayrıca bireyin iş gücünü hafifletmeyi, hastanın tanı ve tedavi ve ayrıca hastalık teşhisinde hekimlerin karar alma süreçlerinde etkili olan bir bilim dalıdır. Belirli insan davranışlarını (örneğin nesneleri alma ve belirli yerlere yerleştirme gibi) ve beşeri düşünme sürecini (örneğin veri hesaplaması, tıbbi teşhis gibi) benzeten sistemler YZ kullanılarak geliştirilebilir. Günümüzde önemli gelişmeler sağlanmış olsa da, YZ alanında yapılan araştırmalar çok az YZ araştırmacıları sürekli olarak yeni teknoloji üretmektedir. Bulanır mantık, YSA, genetik algoritmalar ve uzman sistemler, YZ'nin ana yöntemlerindendir. Bilgisayar sistemleri bir uzman sistem, tasarım, planlama, teşhis, yorumlama, özetleme, genelleme, kontrol etme ve tavsiyelerde bulunma yeteneğine sahiptir (48).

Radyolojik görüntüler kullanılarak yapılan klinik değerlendirmelerin sayısındaki artışın bir sonucu olarak radyologların iş yükünün arttığı bir gerçektir. Bu, eğitilmiş ve deneyimli radyologların eksikliği nedeniyle YZ'nin bu alanda önemli bir rol oynadığını göstermektedir. Literatürde, acil durumları uyarmak, patolojileri görselleştirmek, daha hızlı tanı sağlamak ve kritik insan gücü eksikliği sorunlarına yardımcı olmak için görüntü işleme ve bilgisayarlı görme algoritmalarının geliştirilmesi üzerine çalışmalar hız kazanmıştır (49).

Radyolojik veriler kullanılarak bazı hastalıkların tespit edilebildiğini DÖ yöntemleri ile tahmin edildi. Örneğin göğüs radyografisi dünyada en yaygın radyolojik inceleme türüdür ve çok büyük bir veri kümesine sahiptir. DÖ modellerinin göğüs radyografisinde pnömoni ve pnömotoraks gibi klinik olarak önemli anormallikleri tespit edebildiğini görünmektedir (50). YZ modelleri sayesinde Radyologlar ve doktorlar acil tıbbi müdahale gerektiren hastalarda, YZ algoritmasını erken teşhis ve tedavi için kullanabilir. Ayrıca, görüntüleme sonuçlarına ihtiyaç duydukları aciliyet derecesine göre

öncelik verilerek radyoloji iş akışı optimize edilir. MR anjiyografi kullanarak beyin anevrizması tespiti, karaciğer kitle sınıflandırması, karaciğerde fibroz evrelemesi ve glioma genomik durum tespiti, gibi araştırmalar yapılmaktadır (51). Ayrıca YZ'nin kullanımı, hasta bakımında büyük ilerlemelerin yanı sıra tıbbi tedavi masraflarında da tasarruf sağlama potansiyeline sahiptir. Sağlık sektörü, maliyetleri düşürürken etkinliği ve verimliliği artırmak için yenilikçi fikirlere ihtiyaç duyması ve YZ, hasta teşhisi ve tedavisi ile ilgili cevaplara ilişkin önemli bir faktördür. Başta YZ ve robotik teknolojideki hızlı ilerlemeler, sağlık çalışanları için önemli bir yol gösterici unsurdur (52).

2.6. Makine Öğrenmesi

ML algoritması, yüksek hacimli veri setlerinin işlenmesinde uzmanlık kazanarak bir veri setinin performansını artırmaya yönelik yöntemleri araştıran YZ'nin bir alt koludur. ML veri setinin çıktı değişkenine göre Denetimli, Denetimsiz, Yarı denetimli ve Pekiştirmeli öğrenme yöntemleri ile veri setini açıklar. Girdi verileri ve çıktı verileri arasındaki korelasyonları hakkında araştırmacıya geniş yelpazede bilgi sunar. ML algoritmaları tarafından çözülebilecek çeşitli algısal görevler vardır. ML alanında birincil amaç, verilere dayalı yargılarda bulunup kararlar alabilen, doğru ve tutarlı tahminler üretebilen öngörücü modeller oluşturmaktır. Bu modeller görüntü tanımlama, modelin doğruluğu, doğal dil işleme ve dolandırıcılık tespitinin yanı sıra çok çeşitli uygulamalarda da kullanılabilir. Büyük miktarda verinin varlığı ve bilgi işleme gücündeki ilerlemeler sayesinde, son zamanlarda popüleritesinin artmasına katkıda bulunmuştur (53, 54).

Derin Öğrenme (DÖ), ML'nin bir alt kolu ve bilgisayarların deneyimlerinden öğrenmelerini ve veri setini anlamalarını sağlayan bir kavramlar hiyerarşisini oluşturulmasını sağlayan bir tür YZ dir. Bu sistemin temeli, insan beyninde bulunan biyolojik sinir ağlarından sonra modellenen matematiksel modeller olan yapay sinir ağlarını oluşturmaktadır. DÖ, verilerin bir dizi katman aracılığıyla dönüştürülmesini içeren bir makine öğrenimi türüdür ve sonraki her katman, kendisinden önce gelen katmanın çıktısını girdi modelleri olarak kullanır yani herhangi bir katmanın sonucu bir sonra ki oluşacak katman için girdi verisi olarak işlem görür. DÖ yazılımı, tek ve çok katmanlı işlemler uygulayarak, büyük hacimlerdeki verileri öğrenmek için insan beyninde meydana gelen nöronal aktiviteyi simüle etmeye çalışan bir yapı ile çalışmaktadır (55). Bu tez çalışmasında model seçimi ve eğitimi, değişken seçme, veriyi %80 eğitim ve %20 test verisi olarak bölerek ML modelleri uygulanmıştır, ayrıca tıbbi

görüntülerden ve klinik verilerde elde edilen değişkenler kullanarak EGFR ve KRAS mutasyonlarını tahmin etmek için sınıflandırma performanslarına ilişkin metrikler elde edilmiştir.



3. MATERİYAL VE METOT

3.1. Veri Seti

Bu veri seti, KHDAK olan 211 hastanın kapsamlı tıbbi kayıtlarını ve görüntü verileri bu veri setinde sunulmaktadır. Veri seti, hem tıbbi görüntülemelerde görünen kötü huylu tümörlerin iyi gözden geçirilmiş tanımlamalarını, hem de BT taramalarından elde edilen ilgili görüntülere yönelik niceliksel veriler sunuldu. BT taramaları ve PET/BT taramalarındaki tümörlerin segmentasyon görüntüleri de gözlemlendi. (56). Veri seti 135 (%64) erkek ve 76 (%36) kadın olmak üzere toplam 211 örneklemden oluşacak şekilde derlendi. EGRF mutasyonuna sahip 19 (%14.10) erkek hasta gözlemlendi. EGRF mutasyon olmayan erkek hastalar ise 116 (%85.90) olarak hesaplandı ve KRAS mutasyonu olan erkek hastalar 27 (%20) ve KRAS mutasyonu olmayan erkek 108 (%80) olarak hesaplandı. EGRF mutasyonu olan kadın hastalar 24 (%31.60) olarak hesaplandı. EGRF mutasyonu olmayan kadın hastalar ise 52 (%68,40) kişi hesaplandı. KRAS mutasyonu olan kadın hastalar 11 (%14.50) ve KRAS mutasyonu olmayan kadın hastalar ise 65(%85.50) olarak hesaplandı.

3.2. Klinik Verilerde Ön İşleme Yaklaşımları

Veri madenciliği ve modelleme, veri sürecinin daha sorunsuz ve doğru bir şekilde çalışmasını sağlamak için veri hazırlama ve veri kalitesini iyileştirme sürecidir. Veri temizleme, veri dönüştürme, veri normalleştirme, veri azaltma, RF ile değişken seçme ve diğer süreçler, tipik olarak bu sürece dâhil edildi. Verilerin daha iyi analiz edilebilmesi için verilerin hazır hale getirildi. Bu aşamalar neticesinde veri boşlukları doldurulur, tutarsızlıklar giderilir ve verilerden sonuç çıkarma işlemi tamamlandı (57). Bu tez çalışmasında aşağıdaki veri ön işleme aşamaları verildi.

3.2.1. Veri Temizleme

Veri temizleme, verilerin ön işleminden geçirilmesinin ilk ve en kritik aşamasıdır. Büyük hacimli veri setleri üzerinde gerçekleştirilen veri temizleme süreci, yalnızca ilgili verilerin temizlenmesini sağlar. Bu süreçte, veriler gürültülü ve eksik veriler olarak kategorize edilerek, analiz için uygun hale getirilmesi sağlandı. Veri temizleme, veri

kalitesini artırmak ve güvenilir analiz sonuçları elde etmek için gerekli adımlardan biridir (58).

3.2.2. Veri Dönüştürme

Veri dönüştürme süreci, verileri bir formattan başka bir formata çevirerek, veri bütünlüğünü koruma ve okunabilirliğini artırma amacı taşımaktadır. Bu süreç, verilerin bir sistemden diğerine taşınması, işlenmesi veya saklanması sırasında veri kaynakları arasındaki farklılıkların giderilmesine yardımcı olduğu görüldü. Örneğin, raporlama, iş zekâsı, veri madenciliği, analiz ve veri entegrasyonu gibi alanlarda sıklıkla kullanılır. Veri dönüştürme manuel olarak yapılabileceği gibi otomatikleştirilmiş yöntemlerle de gerçekleştirilebilir, bu da hataları azaltarak daha hızlı ve doğru sonuçlar elde edilmesini sağlar. Normalleştirme, veri tabanında veriyi düzenleme ve işlenebilir yapıya getirme sürecidir. Veri tabanı oluşturma aşamasında benzer veriler farklı ortamlarda tekrar görülme olasılığına sahiptir. Bu durum, veri ekleme, silme ve güncelleme gibi durumların ortaya çıkmasına yol açacaktır. Normalleştirme işlemi ise, verilerin tekrarlama işlemini önleyerek veri tabanında daha az yer kaplaması, etkili, tutarlı verileri bir arada tutması sağlandı (59). İlgili tez çalışmasında veri dönüşümünü VGG16 modeli daha iyi performans sağladığından dolayı, kategorik veriler için scikit-learn (sklearn), normalizasyon için “tensorflow” ve ” keras” kütüphaneleri kullanıldı.

3.2.3. Veri Entegrasyonu

Veri entegrasyonu, farklı veritabanları arasında verilerin uyumlu ve bir araya getirilmesi gereken süreçtir. Genellikle bu süreç, çeşitli veritabanları, uygulamalar ve sistemlerden gelen veriyi içerir, bu verilerin formatları ve yapıları birbirinden farklıdır. Veri raporlaması, verilerin birleştirilmesi adımıyla anlamlı sonuçlar elde etmeye çalışıldı. Veri entegrasyon süreci, verileri tek veri seti haline getirmesi için genel olarak sırasıyla; verinin toplanması, temizlenmesi, dönüştürme yönteminin uygulanması ve verilerin birleştirilmesi adımlarını içermektedir (60). Bu tez çalışmasında, “pandas” kütüphanesi kullanılarak görüntülerden elde edilen veriler bir dataframe kaydedilip klinik veriler ile birleştirildi.

3.2.4. Değişken Seçme Yöntemi

Makine Öğrenmesi (ML) yönteminin uygulama aşamasında değişken seçme, veride gereksiz, tekrar eden veya veri setinde düşük etki gösteren özellikler varsa bunları belirleyip veri setinden çıkarma sürecini ifade eder. Bu süreç elde edilen modelin hızlı ve daha doğru sınıflandırma sonucu üretmesine imkân sağlar (61). Değişken Filtreleme, Veri setinde değişkenleri önemlilik düzeylerine göre seçme imkânı sağlamak için istatistik ve bilgi tabanlı yöntemler aracılığıyla modelleri eğitime aşamasında bağımsız bir şekilde çalışır. En çok kullanılan filtre yöntemleri arasında Pearson Korelasyon Katsayısı, ANOVA (Analysis of Variance), Kendall ve Spearman Korelasyonlar, Karşılıklı Bilgi (Mutual Information) yöntemleri sayılabilir (62, 63). Gömülü Yöntemler ise, değişken seçimi, gömülü yaklaşımlar kullanılarak ML eğitime entegre edildi. Bu teknikler otomatik olarak en önemli değişkenleri seçer ve model eğitim sürecinde gereksiz olanlar elimine edildi. Gömülü teknikler, dallanma süreçlerinde özelliğin öneminden yararlanan Lasso Regresyon, Karar ağaçları, Rastgele Orman ve Gradyan Destekli Ağaçlar gibi yaygın olarak kullanılan değişken seçimi algoritmalarıdır (64).

3.2.5. Boyut İndirgeme

Büyük veri setlerinin yönetilmesini ve analiz edilmesini kolaylaştıran temel veri hazırlama tekniklerinden biri boyut indirgemedir. Bu yöntem, yüksek boyutlu veri kümelerindeki gereksiz veya düşük bilgi içeren özellikleri ortadan kaldırarak veri miktarını azaltmayı amaçlar. Veri boyutunun azaltılması özellikle veri analitiği, ML ve görselleştirme gibi alanlarda çok önemlidir. Bu yöntem, veri setinin temel yapısından ödün vermeden gereksiz özellikleri kaldırarak model performansını artırır ve verilerin yorumlanmasını kolaylaştırır (65). Bu tez çalışmasında, “tensorflow”, “keras” ve “vgg16” kütüphaneler kullanarak VGG16 modelinin son katmanından çıkarılan değişkenler, görüntülerin yüksek boyutlu olduğu gözlemlendi. Bu, orijinal görüntü boyutuyla karşılaştırıldığında boyutta önemli bir azalma görüldü.

3.2.6. Veri Bölme

Bir veri kümesini farklı alt kümelere bölme tekniği veri bölümlenmesi olarak bilinir ve ML modeli eğitimi ve değerlendirmesini geliştirmek için kullanılır. Bu işlem, modelin gerçek veriler üzerindeki performansını daha doğru bir şekilde değerlendirmek için önem taşımaktadır. Eğitim ve test kümelerinin yanı sıra çapraz doğrulama, verileri

ayırmak için kullanılan popüler tekniklerdir. Veri setinin tipik bir bölümü %20-30 test verisi ve %70-80 eğitim verisi olarak belirlenir. Test seti, modelin doğruluğunu değerlendirmek için kullanılırken, eğitim seti, modeli eğitmek için kullanıldı. Veri seti, k-kat çapraz doğrulama gibi teknikler kullanılarak daha küçük alt gruplara ayrılır (66). Bu tez çalışmasında, veri seti rasgele olarak %80 eğitim ve %20 test olacak şekilde ikiye bölündü. Model performanslarının doğru değerlendirilmesi için 5 katlı çapraz geçerlilik kullanıldı.

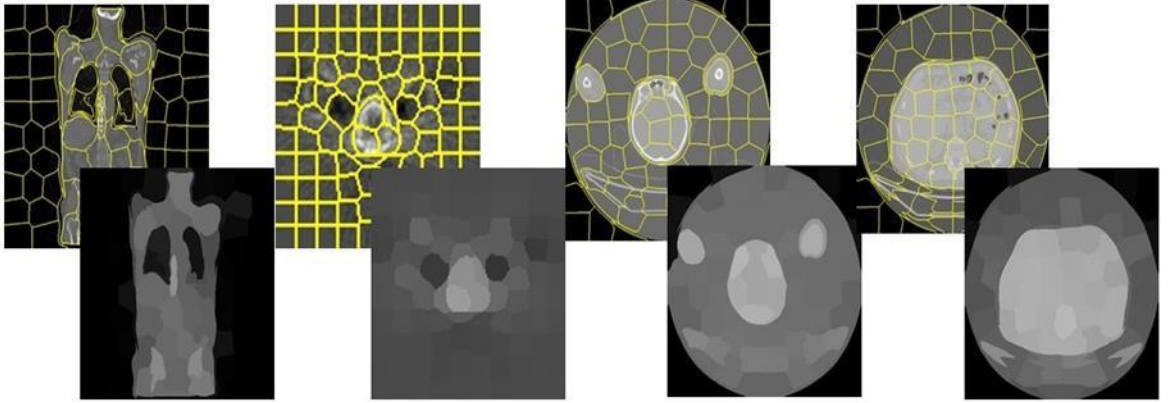
3.3. Görüntü Verilerinde Ön İşleme Yaklaşımları

Görüntü ön işleme, dijital görüntülerin işlenmemiş verilerini analiz edilebilir ve anlamlı hale getirmek için kullanılan bir dizi teknik ve yöntemler bütünüdür. Bu işlem, görüntülerin kalitesini artırmak, gürültüyü azaltmak ve belirli özellikleri vurgulamak amacıyla çeşitli adımlardan oluşur. Görüntüdeki parlaklık, kontrast ve renk dengesini ayarlamak, ilgisiz bölgeleri kırmak için görüntünün kesilmesi, görüntünün boyutunu değiştirmek için yeniden boyutlandırma, görüntüdeki rastgele piksel gürültüsünü azaltmak için filtrelerin kullanılması, görüntüdeki nesnelerin sınırlarını belirlemek için kenar algılama yöntemlerinin kullanılması ve görüntünün kontrastını artırmak için histogram eşitleme işleminin yapılması gibi işlemler uygulanarak görüntü analiz edilebilir duruma getirilir (67). Bu tez çalışmasında kullanılan görüntü ön işleme aşamaları kapsamlı bir şekilde aşağıda açıklandı.

Görüntü normalizasyonu, dijital parçacıkların piksel dağılımlarını belirli bir aralık veya dağılıma dönüştürme sürecidir. Görüntü verilerinin işlenmesi ve analizi bu işlemle daha kolay hale getirildi. ML ve bilgisayarla bakım uygulamaları, özellikle desen gösterme, segmentasyon ve nesne tanımda normalizasyon süreci için çok önemlidir. Görüntü normalizasyonu, kontrast ve aydınlatma farklılıklarını modellemeye öğrenme dağılımı için kullanıldı. Görüntülerin dağılımlarının farklı olması, farklı aydınlatma koşullarının bir sonucudur. Bu farklılıklar göz önünde bulundurularak normalizasyon, modellerin daha kapsamlı ve yaygın sonuçların üretilmesine imkân tanır (68). Bu tez çalışması kapsamında görüntüyü yeniden boyutlandırma işleminde “cv2 (OpenCV)” kütüphanesi kullanıldı. Farklı boyutlara sahip görüntüleri tutarlı işlenmesi için görüntülere ilişkin boyutlar sabit bir boyut kullanıldı. Ayrıca, “cv2 (OpenCV)” kütüphanesi kullanılarak kontrast sınırlı uyarlanabilir histogram eşitleme (CLAHE:

Contrast Limited Adaptive Histogram Equalization) görüntü ayrıntılarının görünürlüğü iyileştirmek için kullanıldı.

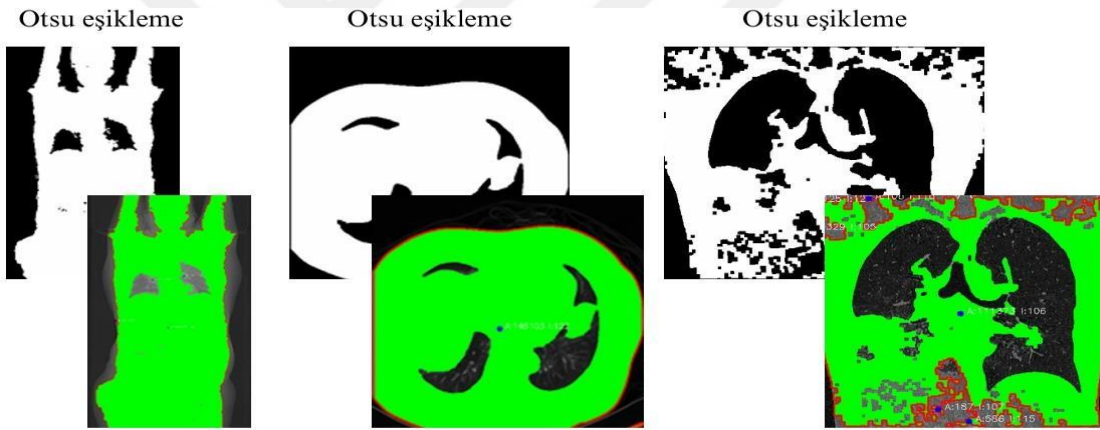
Görüntü Bölütleme (Segmentasyon), görüntüyü, renk, doku, şekil ve parlaklık gibi doğal özelliklerine göre birbirinden anlamlı parçalara veya nesnelere ayırmak görüntü bölütlenmesi olarak ifade edilir. Görüntüdeki farklı nesnelerin veya bölgelerin daha iyi anlaşılması ve analiz edilmesi için kullanılan bir tekniktir. Tıbbi görüntülemelerde tümörlerin tespiti, uydu görüntülerinde arazi sınıflandırması veya otonom araçlarda yol ve engelin algılanması gibi uygulamalarda yaygın olarak kullanılan bir yöntemdir. Ayırma, her bir bölgenin ayırt edici özelliklerini koruyarak bir görüntünün daha derin incelenmesine izin verir. İlgili tez çalışmasında Otsu eşikleme ve SLIC Süper Piksel Segmentasyonu için “scikit-image” kütüphanesi kullanıldı. Morfolojik İşlemler (gürültüyü azaltma ve görüntü sınırlarını belirleme) ve bağlantılı pikseller, tümör bölgelerini belirlemek için Bağlantılı Bileşen Etiketleme için “SciPy (scipy.ndimage)” kütüphanesi kullanıldı. Kontur bulma işlemi için “cv2(OpenCV)” kütüphanesi kullanıldı. Bu tez çalışmasına ilişkin elde edilen görüntüler aşağıda sunulmuştur; her sütunun üst sırasında sarı çizgiler süper piksel sınırlarını ve alt sırada ise segmentasyon sonuçları gösterilmektedir (Şekil 3.1).



Şekil 3.1. BT görüntülerinde süper piksel segmentasyonu ve sonuçları

Görüntü Eşikleme (Thresholding), Bir tasarımı birden fazla görüntü grubuna ayırır. Her bir pikselin değerlerinin belirli bir kritik değere eşit veya daha büyük olup olmadığı belirlenir. Eşikleme yaklaşımı, özellikle resim segmentasyonu ve alanlar arası nesne tespitinde sıklıkla kullanılır. Eşikleme yöntemlerini basit ve otsu eşikleme yöntemleri olarak iki grupta inceleyecek olursak, basit eşikleme ile pikselleri iki sınıfa ayırmak için sadece bir eşik kullanılır. Örneğin, gri tonlamalı bir görüntüde, bir pikselin

değeri belirlenen eşik değerinden küçükse siyah, büyükse beyaz olarak atanır. Basit eşikleme yöntemini kullanırken atılacak adımlar şunlardır: Öncelikle bir kesme noktası belirlendi. İlgili eşik değeri görüntüdeki her bir pikseli karşılaştırmak için kullanıldı. Bir piksel sınıfı, eşik değerinden büyük veya eşit durumlarında bir sınıfta değerlendirilir. (71). Otsu eşikleme ise, histogramlara dayalı bir yaklaşım kullanarak görüntüleri otomatik bir şekilde iki sınıfa ayırmak için ideal eşik değerini hesaplar. Otsu eşikleme uygulandığında bir görüntü histogram grafiği oluşturuldu. Histogramı iki sınıfa ayırmak için en etkili eşik değeri belirlendikten sonra uygun eşik değeri kullanılarak görüntü iki sınıfa ayrıldı (72). Çalışmada elde edilen görüntüler dikkate alındığında, Otsu eşikleme işlemede siyah beyaz olarak meydana gelidi ve ikili maskelendi. İlgilenilen bölge beyaz olarak belirlendi ve bu bölge tümörün kaba segmentasyonunu oluşturdu. Eşikleme sonrası yeşil olarak belirlenen bölgeler normal doku, gri belirtiler tümör bölgesi ve mavi ile belirtilen noktalar ise referans olarak alınacak noktalar olarak belirlendi (Şekil 3.2).



Şekil 3.2. Otsu eşikleme sonrası görüntünün maskelenmesi ve referans noktalarının belirlenmesi

3.4. Evrişimsel Sinir Ağları

Evrişimsel sinir ağları (Convolutional Neural Networks (CNN)) esasen tamamen bağlantılı, havuzlama ve konvolüsyonel katmanlardan oluşan bir mimaridir (73). Mimarideki ardışık katmanlar çeşitli değişkenleri çıkarmak için kullanıldıktan sonra, sınıflandırıcı katman olan son katman sınıflandırmayı gerçekleştirmek için kullanıldı. Belirlenen sınıflar ile hata olarak bilinen gerçek sınıflar arasındaki tutarsızlık, çıkarılan özellikler kullanıldı. Geri yayılım tekniği kullanılarak ağırlık değerleri sürekli olarak güncellenir, böylece değişken çıkarma süreci iyileştirilirken ortaya çıkan hata miktarı da

en aza indirildi. Görüntü sınıflandırma için sıklıkla kullanılan CNN mimarisinde farklılıklar ve eklemeler yapılarak farklı CNN tabanlı tasarımlar oluşturulmuştur. Bu modifikasyon ve eklemelerin amacı CNN mimarisinin performansını artırma düşüncesidir (74). Bu tez çalışmasında ise, 13 evrişim katmanı ve 3 tam bağlantılı katman olmak üzere toplamda 16 katmandan oluşan VGG16 yöntemi ile, görüntüden kenar, doku, şekil gibi değişkenler seçilerek analize dahil edildi.

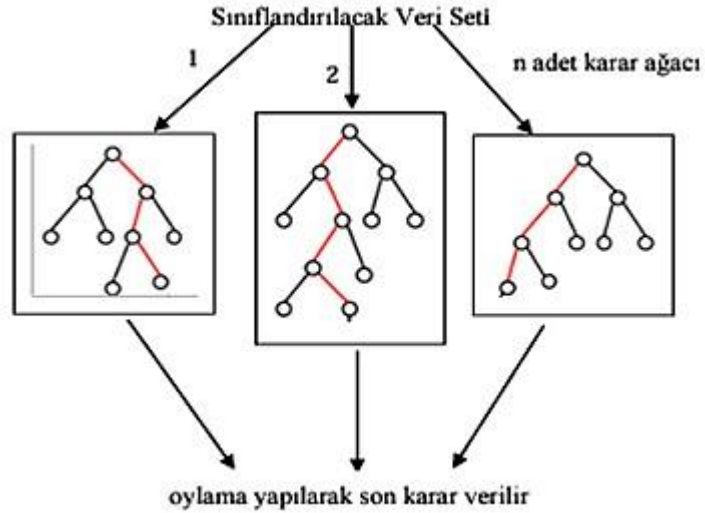
3.4.1. Random Forest (RF)

RF yönteminin geliştirilmesinde Leo Breiman ve Adele Cutler sınıflandırma ve regresyon sorunlarının çözümünde RF algoritması geliştirildi. Bu algoritma, çok sayıda karar ağacını entegre ederek ve nihai sonuca ulaşmak için “voting” olarak bilinen oylama tekniğinden faydalanır. Bu yöntem ile birbirinden bağımsız olarak, her ağaç eğitime tabi tutulur ve bir olasılık değeri üretilir. Sınıflandırma söz konusu olduğunda, her bir ağaç tarafından önerilen sınıflar sayılır ve en çok oyu alan sınıf, modelin tahmini olarak seçilir. Ağaçlar tarafından kestirilen sayısal değerlerin regresyon durumlarının ortalaması hesaplandı. Yeniden örneklemenin bootstrap yöntemi, veri kümesinin çeşitli alt kümelerini üretmek için kullanılır ve bunlar daha sonra karar ağaçlarını bağımsız bir şekilde eğitmek için kullanılır. Bu tekniğin kullanılmasıyla modelin genelleme oranı artar, hata oranı azaltılır ve ağaçlar arasındaki çeşitlilik arttırılmış olur (75). RF yöntemi, görüntü sınıflandırma işlemlerinde yoğun olarak kullanıldığı pek çok çalışma alanından biridir. RF'nin çok sayıda özelliğın olduğu veri setlerinde başarılı bir performansa sahip olduğu, güçlü gürültü direnci gösterdiği ve yüksek sınıflandırma doğruluğu elde ettiğı bilinmektedir. Nihai sınıflandırma sonucunda, veriler üzerinde eğitilen birçok ağacın bulgularını bütünleştiren RF modelindeki bu karar ağaçlarının çoğunluk oyuna göre istatistiksel kararlar verilir (76).

RF algoritmasını Extreme Gradient Boosting (XGBoost), Destek vektör makineleri (DVM), CatBoost gibi algoritmalarından ayıran en önemli özellik bootstrap yönteminin kullanılmasıdır. Bu yöntem, birden fazla karar ağacı algoritmasını eğitme yeteneğine sahiptir. Nihai sonuca ulaşmak için ilgili veri setinden elde edilen modeller birleştirildi. Girdi veri değişkeninin model üzerindeki etkisini azaltmak için bootstrap tekniğinden yararlanır. Bootstrap, test verileri için sınıflandırma hatasını azaltarak, modelin mevcut verilerle eşleşmesini ve genelleştirilmiş sonuçlar üretmesini sağlar. Ağaç sayısı, RF yapısının en önemli unsurlarından biridir ve modelin performansı üzerinde

doğrudan bir etkiye sahip olduğundan, modelin performansını şekillendiren kritik bir bileşendir. Ağaç sayısı, test verileri üzerindeki hata oranını sınırlar. Bu, modelin aşırı derecede uyumlu olmasını önler.

RF algoritmasının güvenilirliğini ve kararlılığını artırdığı için, daha az eğitilir. Sonuç olarak, algoritma hem genelleme hem de modelin doğruluğu açısından oldukça güçlüdür ve bu nedenle tahmin ve sınıflandırma sorunları için güvenilir bir çözüm sağlandı (77). Veri kümesini rastgele parçalara böldükten sonra her bir parça için bir karar ağacı oluşturan RF algoritmasına ilişkin karar ağaç yapısı sunuldu (78). Bu tez çalışmasında değişken seçime yöntemi için “scikit-learn” kütüphanesi, RandomForestClassifier sınıfı ve feature_importances_ özelliği kullanılarak değişken seçimi yapıldı. RF algoritması şematik gösterimi sunuldu (Şekil 3.3).



Şekil 3.3. Random Forest algoritması karar ağaç yapısı

3.4.2. Aşırı Gradyan Artırma Algoritması

Aşırı Gradyan Artırma (XGBoost:Extreme Gradient Boosting), Chen ve Guestrin tarafından geliştirilen regresyon ve sınıflandırma yöntemlerinde kullanılan güçlü bir ML algoritmasıdır. Özellikle yapılandırılmış verilerde üstün performans göstermektedir. XGBoost, gradyan artırma algoritmasının optimize edilmiş bir versiyonudur. Bu algoritma, modelin hatalarını düzeltmek için zayıf öğrencileri sıralı olarak birleştirerek çalışır ve bu süreçleri daha verimli hale getirildiği gözlemlendi. XGBoost, sağlık, finans,

gibi çeşitli sektörlerde ve klinik görüntülerde yaygın olarak kullanılır. Büyük veri setlerinde ve eksik hücrelere sahip verilerde etkili performans göstermektedir. (79). Yapay sinir ağları (YSA) veya DVM ile karşılaştırıldığında, XGBoost hiperparametrelerin daha basit ayarlanmasını mümkün kılar ve aykırı değerlere karşı daha dirençlidir. XGBoost, Gradient Boosting algoritması üzerinde geliştirilmiş bir ML algoritmasıdır. Matematiksel bir ifade olarak XGBoost'un teorik temeli şu şekilde ifade edilir (80):

$$\hat{y}_i = \sum_{j=1}^M f_j(x_i) \quad (1)$$

$$\hat{y}_i: \text{ö} \rightarrow \text{ö}$$

$$M: \text{ö} \rightarrow \text{ö}$$

$$\text{ö} \rightarrow \text{ö}$$

Her yeni bir adımda, XGBoost algoritması, önceki model tarafından ortaya çıkan ve artık hatalar olarak da bilinen tahmin hatalarını ortadan kaldırmak için bir regresyon ağacı üretildi (81). Bu çalışmada, XGBoost algoritması için “XGBClassifier” kütüphanesi ve 200 yineleme ile modellere ilişkin sınıflandırma performansları elde edildi.

3.4.3. Categorical Boosting Algoritması (CatBoost)

CatBoost tekniği hem sayısal hem de kategorik verilerde çalışabilen ML yöntemlerindendir. Düşük frekanslı özelliklere ve yüksek yoğunluğa sahip alanlara algoritmanın bir önceki değerleri ekleyerek, CatBoost gürültüdeki noktaları yönetir ve bu algoritmanın ayırt edici özelliklerinden biri olan aşırı uyumu en aza indirir. Gradyan destekli karar ağaçları CatBoost yönteminin geliştirilmesinin temelini oluşturmaktadır. CatBoost, gradyan inişi yaklaşımının ürettiği tahmin değerlerindeki sapmanın üstesinden gelerek verileri doğru bir şekilde kavrar ve sonuçları değerlendirir. (82, 83). Bu çalışmada, CatBoost algoritması için “CatBoostClassifier” kütüphanesi ve 200 yineleme ile sınıflandırma performansına ilişkin metrikleri elde etmede kullanıldı.

3.4.4. Destek Vektör Makineleri

Destek Vektör Makineleri (DVM) algoritması, geleneksel örüntü tanıma yöntemlerinin, eğitim seti üzerindeki performansı optimize ederek ampirik hata olasılığını azaltmak için tasarlanmış bir algoritmadır. Öte yandan, DVM öncelikle sabit ancak bilinmeyen bir veri olasılık dağılımına dayalı olarak henüz gözlemlenmemiş örüntüleri yanlış kategorize etme tehlikesini ifade eden yapısal riski en aza indirmekle ilgilenir. Olasılıkta tekdüze yakınsama fikrine göre, genelleme hatası üzerindeki bir üst kısıtlamayı en aza indirmeye eşit olan bu yeni tümevarım ilkesi geliştirildi (84). Bu çalışmada, CatBoost algoritması ile model sınıflandırma performansı için “scikit-learn” kütüphanesi, model için “Support Vector Classifier” ve çekirdek model için “Radial Basis Function” kullanıldı ayrıca `class_weight='balanced'` sınıf dengesizliği ele alındı.

3.5. Biyoistatistiksel Analizler

Bu tez çalışması kapsamında, nicel veriler ortalama $\pm$ standart sapma ile özetlenirken, nitel veriler ise sayı (yüzde) ile özetlendi. Veri analizinden sonuç çıkarımından önce veri setinde var olan eksik/kayıp veya aşırı uç değerlerin belirlenmesine yönelik analizler yapıldı. Uygun yöntemler ile veri setinde düzeltmeler yapıldı. Normal dağılım varsayımına uygunluğun test edilmesi Shapiro-Wilk testi ile gerçekleştirildi ve böylece veri analizinde kullanılan hipotez testleri belirlendi. İstatistiksel anlamlılık düzeyi $p < 0.05$ olarak kabul edildi.

Tez çalışması kapsamında yapılan istatistiksel analizlerin gerçekleştirilmesinde IBM SPSS Statistics version 27.0 paket programı kullanıldı (85). Biyomedikal veri ve görüntü analizlerinin sınıflandırılmasında kullanılan ML yöntemleri için Python programlama dili ile “tensorflow, keras, preprocessing.image”, “seaborn”, “pandas”, “OpenCv”, “traceback”, “os” “pydicom”, “sklearn.ensemble”, “StratifiedKFold” “tqdm”, “sklearn.metrics”, “tensorflow, keras”, “torch”, “skimage.segmentation”, “glob” kütüphaneleri kullanıldı.

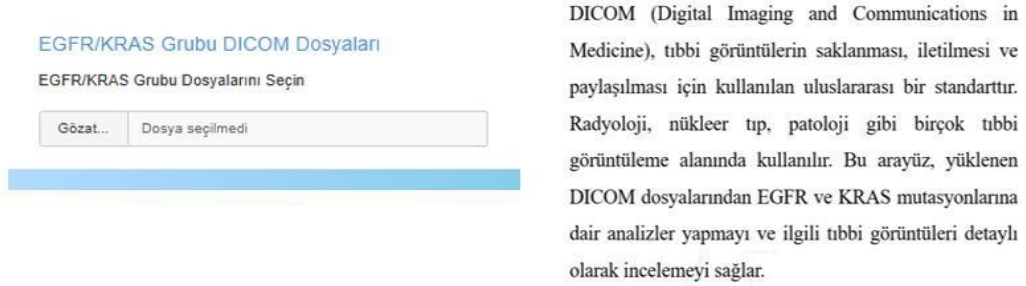
3.6. Geliştirilen Web Tabanlı Yazılım

Bu tez çalışmasında, KHDAK’lı hastaların biyomedikal görüntü verileri ile klinik verilerinin işlenmesi, sınıflandırılması görevlerini yerine getirmesi için geliştirilen web tabanlı yazılım “Shiny” kütüphanesi kullanılarak tasarlandı. Ayrıca arayüz

oluşturulmasında “pydicom”, “OpenCv”, “numpy”, “tensorflow, keras”, “pandas” paketleri de kullanıldı. İlgili arayüz, “Ana Sayfa”, “Görüntü Analizi ve Sınıflandırma” sekmelerinden oluşturuldu.

3.6.1. Ana Sayfa Sekmesi

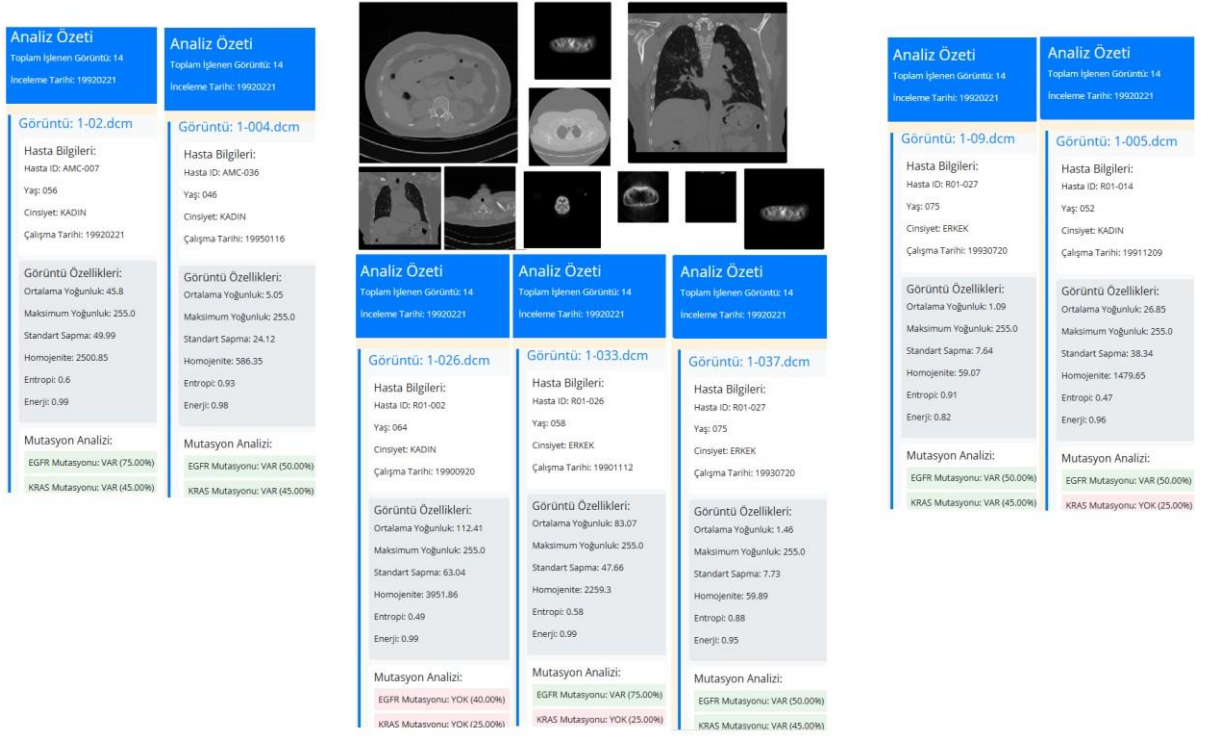
Görüntü verilerinin sınıflandırılması için geliştirilen yazılımda DICOM görüntüleri aşağıdaki buton ile içe aktarıldı (Şekil 3.4).



Şekil 3. 4. Ana sayfa sekmesi

3.6.2. DICOM Görüntülerinin Okunması

Söz konusu görüntülerin yüklenmesi ile görüntü ön işleme işlemleri; değişken seçme, görüntü tipini kontrol ederek normalize işleminin gerçekleştirilmesi, Görüntü azaltma, Kontrast iyileştirme, Keskinleştirme ve Medyan filtre uygulanarak “.dcm” uzantılı DICOM görüntü dosyalarına ilişkin meta veriler (Hasta bilgileri) elde edildi. DICOM dosyaları görüntü ön işleme adımları uygulanarak “.jpg” formatına dönüştürülüp hasta bilgileri ve mutasyonun varlığı veya yokluğu hakkında kullanıcıya bilgi verilmesi sağlandı. Görüntü verilerine ilişkin bilgiler aşağıda sunuldu (Şekil 3.5).



Şekil 3. 5. DICOM görüntülerinden meta verilerin çıkarılması

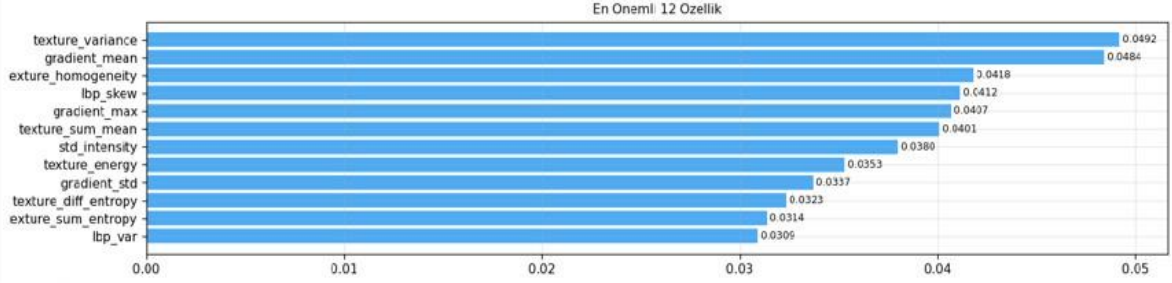
3.6.3. Görüntü Analizi ve Sınıflandırma Sekmesi

Model analizi ve sınıflandırma sekmesinde Dicom görüntü verileri .jpg formatına dönüştürüldükten sonra görüntülere ilişkin en önemli 12 değişken seçildi. Seçilen bu değişkenler ile ML modellerinden RF, XGBoost, CatBoost, SVM modellerinin performans metrikleri hesaplandı (Şekil 3.6).

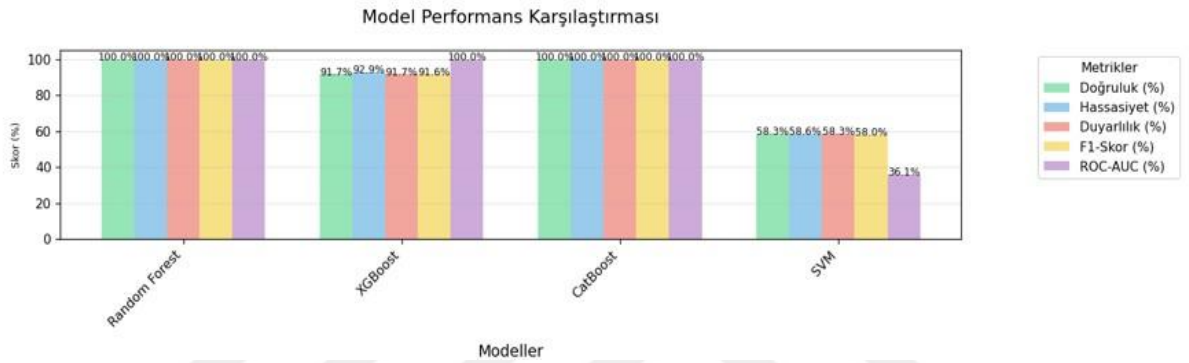
Model Performans Metrikleri

	Doğruluk (%)	Hassasiyet (%)	Duyarlılık (%)	F1-Skor (%)	Dengelenmiş Doğruluk (%)	ROC-AUC (%)
Random Forest	100.00	100.00	100.00	100.00	100.00	100.00
XGBoost	91.67	92.86	91.67	91.61	91.67	100.00
CatBoost	100.00	100.00	100.00	100.00	100.00	100.00
SVM	58.33	58.57	58.33	58.04	58.33	36.11

Özellik Önemlilikleri



Model Karşılaştırması



Şekil 3. 6. Modellere ilişkin çıktılar

4. BULGULAR

Bu tez çalışmasında kullanılan veri setinde, Epidermal Büyüme Faktörü Reseptörü (EGRF) ve Kirsten Sıçan Sarkomu viral onkojenik homolog (KRAS) mutasyonların varlığı ve yokluğu hakkında bilgi sunulmaktadır. Veri seti 135 (%64) erkek ve 76 (%36) kadın olmak üzere toplam 211 örnek hesaplandı. EGRF mutasyonuna sahip erkek hastalar 19 (%14,10) olarak hesaplandı. EGRF mutasyon olmayan erkek bireyler ise 116 (%85,90) olarak hesaplandı ve KRAS mutasyonu olan erkek hastalar 27 (%20) ve KRAS mutasyonu olmayan erkek hastalarda ise 108 (%80) olarak elde edildi. EGRF mutasyonu olan kadın hastalar 24 (%31,60) ve EGRF mutasyonu olmayan kadın hastalar ise 52 (%68,40) şeklinde hesaplandı. KRAS mutasyonu olan kadın hastalar 11 (%14,50) ve KRAS mutasyonu olmayan kadın hastalar ise 65 (%85,50) olarak hesaplandı. Kategorik değişkenlere ilişkin demografik özellikler Tablo 4.1’de sunuldu.

Tablo 4.1. Kategorik verilere ilişkin tanımlayıcı istatistikler

Değişkenler	Kategoriler	EGFR Mutasyon		p	KRAS Mutasyon		p
		Var	Yok		Var	Yok	
		N(%)	N(%)		N(%)	N(%)	
Nüks	Hayır	30 (69.80)	127 (75.60)	0.558*	28 (73.70)	129 (74.60)	0.987
	Evet	13 (30.20)	41 (24.40)		10 (26.30)	44 (25.40)	
Patient affiliation	Veterans	8 (18.60)	85 (50.60)	<0.001*	19 (50.00)	74 (42.80)	0.417*
	Affairs Sistemi					0%	
	Stanford Sağlık					99	
	Sistemi/Hastan	35 (81.40)	83 (49.40)		19 (50.00)	57.20	
Cinsiyet	Erkek	19 (44.20)	116 (69.00)	0.002*	27 (71.10)	108 (62.40)	0.414*
	Kadın	24 (55.80)	52 (31.00)		11 (28.90)	65 (37.60)	

Etnik köken	Kafkas	32 (74.40)	140 (83.30)	0.552**	33 (86.80)	139 (80.30)	0.508**
	Yerli Hawaii/Pasifik Adalı	0 (0.00)	3 (1.80)		0 (0.00)	1 (1.70%)	
	Asya	10 (23.30)	14 (8.30)		2 (5.30)	22 (12.70)	
	Afro-Amerikan	0 (0.00)	6 (3.60)		2 (5.30)	4 (2.30)	
	Hispanik/Latin					5	
Sigara içme durumu	o	1 (2.30)	5 (3.00)	<0.001*	1 (2.60)	2 (2.90)	0.119**
	Sigara kullananlar	2 (4.70)	31 (18.50)		8 (21.10)	25 (14.50)	
	Daha önce kullananlar	19 (44.20)	111 (66.10)		26 (68.40)	104 (60.10)	
	Sigara kullanmayanlar	22 (51.20)	26 (15.50)		4 (10.50)	44 (25.40)	
Histoloji	Adenokarsinom	43 (100.00)	129 (76.80)	0.002**	37 (97.40)	135 (78.00)	0.010**
	Skvamöz hücreli karsinom	0 (0.00)	35 (20.80)		0 (0.00)	35 (20.20)	
	T1a	2 (4.70)	38 (22.60)		8 (21.10)	32 (18.50)	
	T1b	8 (18.60)	23 (13.70)		5 (13.20)	26 (15.00)	
	T2a	27 (62.80)	69 (41.10)		16 (42.10)	80 (46.20)	
Patolojik Tümör evresi	T2b	1 (2.30)	9 (5.40)	0.080**	2 (5.30)	8 (4.60)	0.750
	T3	3 (7.00)	18 (10.70)		6 (15.80)	15 (8.70)	
	T4	1 (2.30)	6 (3.60)		0 (0.00)	7 (4.00)	
	Tis	1 (2.30)	5 (3.00)		1 (2.60)	5 (2.90)	
	N0	38 (88.40)	140 (83.30)		32 (84.20)	146 (84.40)	
Patolojik Lenf Düğüm evresi	N1	2 (4.70)	13 (7.70)	0.099	3 (7.90)	12 (6.90)	0.969**
	N2	3 (7.00)	15 (8.90)		3 (7.90)	15 (8.70)	

Patolojik	M0	42 (97.70)	164 (97.60)		36 (94.70)	170 (98.3)	0.221
	Metastaz evresi			0.989**	0	**	
	M1b	1 (2.30)	4 (2.40)		2 (5.30)	3 (1.70)	
	G1	7 (16.30)	25 (14.90)		5 (13.20)	27 (15.6)	
	G2	29 (67.40)	96 (57.10)		21 (55.30)	104 (60.1)	
Histopatolojik					0	0	0.650
Dereceler	G3	0 (0.00)	33 (19.60)	0.014**	9 (23.70)	24 (13.9)	**
	Diğer, Tip I	4 (9.30)	5 (3.00)		1 (2.60)	8 (4.60)	
	Diğer, Tip II	3 (7.00)	9 (5.40)		2 (5.30)	10 (5.80)	
Lenfovasküler invazyon	Var	3 (7.00)	18 (10.70)	0.579*	4 (10.50)	17 (9.80)	0.968
	Yok	40 (93.00)	150 (89.30)		34 (89.50)	156 (90.2)	
Plevral invazyon	Hayır	34 (79.10)	135 (80.40)		30 (78.90)	139 (80.3)	
(elastik, visseral veya parietal)	Evet	9 (20.90)	33 (19.60)	0.998*	8 (21.10)	34 (19.7)	*
	Hayır	37 (86.00)	125 (74.40)		28 (73.70)	134 (77.5)	
Adjuvan Tedavi	Evet	6 (14.00)	43 (25.60)	0.158*	10 (26.30)	39 (22.5)	0.774
	Hayır	37 (86.00)	125 (74.40)		28 (73.70)	134 (77.5)	
Kemoterapi	Evet	6 (14.00)	43 (25.60)	0.158*	10 (26.30)	39 (22.5)	0.774
	Hayır	41 (95.30)	154 (91.70)		35 (92.10)	160 (92.5)	
Radyasyon	Evet	2 (4.70)	14 (8.30)	0.535*	3 (7.90)	13 (7.50)	0.980
	Ölü	7 (16.30)	56 (33.30)		10	53 (30.6)	

Hayatta Kalma				(26.30)	0)	0.741
Durumu				0.046*		*
	Sağ	36 (83.70)	112 (66.70)	28	120 (69.4)	
				(73.70)	0)	

EGFR: Epidermal Growth Factor Receptor (Epidermal Büyüme Faktörü Reseptörü); KRAS: Kirsten Rat Sarcoma viral oncogene homolog (Kirsten Sıçan Sarkomu viral onkogen homologu); G1: İyi Farklılaştırılmış; G2: Orta derecede farklılaştırılmış; G3: Zayıf şekilde farklılaştırılmış; Tip I: İyi ile orta düzeyde farklılaştırılmış; Tip II: Orta ila kötü derecede farklılaşmış

* : Pearson Ki-kare; **: Fisher's Exact Test



EGFR mutasyon durumunu gösteren tablo incelendiğinde, cinsiyet ($p=0.002$), sigara içme durumu ($p<0.001$), histoloji ($p=0.002$), histopatolojik derece ($p=0.014$) ve hastanın hayatta kalma durumu ($p=0.046$) değişkenlerinde EGFR mutasyon durumu (var, yok) bazında istatistiksel olarak anlamlı farklılık bulundu ($p<0.05$). EGFR mutasyonu olan hastalarda Stanford Sağlık Sistemi/Hastanesine bağlı hasta oranı (%81.4), cinsiyet açısından kadın hasta oranı (%55.8) erkek hasta oranından daha olarak hesaplandı (%44.20). Hiç sigara içmemiş hasta oranı (%51.2) hesaplandı. Ayrıca, EGFR mutasyonu olan hastaların tamamı adenokarsinom histolojisine sahip olduğu ve bu hasta grubunda sağkalım oranı (%83.7) olarak hesaplandı. Diğer yandan, patolojik evreleme (T, N, M), lenfovasküler invazyon, plevral invazyon ve adjuvan tedavi parametreleri istatistiksel açıdan anlamlı farklılık bulundu ($p>0.05$). Ayrıca ilgili tablo KRAS mutasyon durumuna göre incelendiğinde ise sadece histoloji değişkeni istatistiksel olarak anlamlı farklılık saptandı ($p<0.05$; $p=0.010$). KRAS mutasyonu olan hastaların %97.4'ü adenokarsinom histolojisine sahip olduğu fakat diğer tüm demografik ve klinikopatolojik değişkenlerde (cinsiyet, etnik köken, sigara içme durumu, patolojik evreleme, histopatolojik derece, lenfovasküler invazyon, plevral invazyon, adjuvan tedavi ve sağkalım) KRAS mutasyon durumuna (var, yok) göre istatistiksel olarak anlamlı farklılık bulunmadı ($p>0.05$)

Tablo 4.2. Nicel verilere ilişkin tanımlayıcı istatistikler

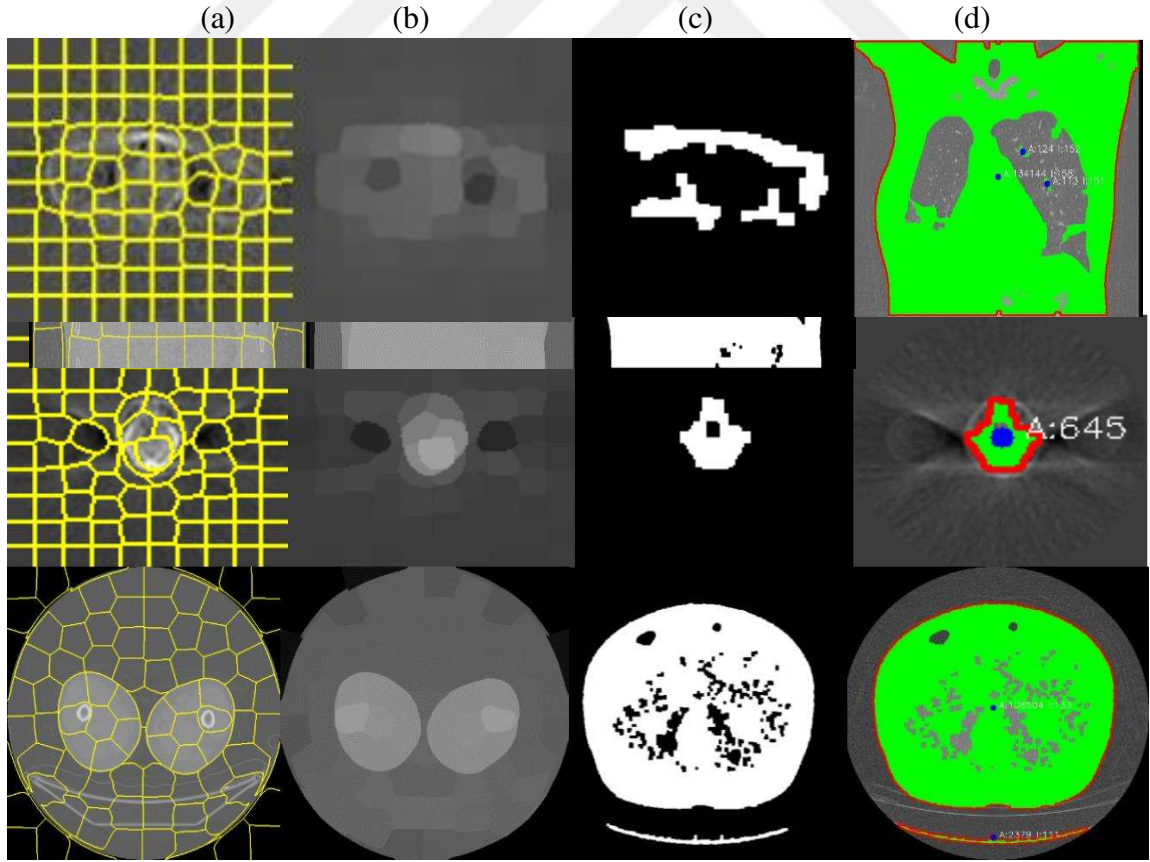
Değişkenler	EGFR Mutasyon Durumu			KRAS Mutasyon Durumu		
	Var	Yok	p^*	Var	Yok	p^*
	Ortalama $\pm$ SS	Ortalama $\pm$ SS		Ortalama $\pm$ SS	Ortalama $\pm$ SS	
Histolojik Tanı Yaşı	68 $\pm$ 10	68 $\pm$ 10	0.924	66 $\pm$ 10	68 $\pm$ 10	0.165
Ağırlık (lbs)	122 $\pm$ 18	123 $\pm$ 22	0.783	128 $\pm$ 26	121 $\pm$ 20	0.096
Pack Years	32 $\pm$ 16	41 $\pm$ 24	0.004	43 $\pm$ 28	38 $\pm$ 21	0.257
BT ve ameliyat arasındaki günler	39 $\pm$ 27	53 $\pm$ 68	0.209	48 $\pm$ 39	51 $\pm$ 67	0.789

EGFR: Epidermal Growth Factor Receptor (Epidermal Büyüme Faktörü Reseptörü); KRAS: Kirsten Rat Sarcoma viral oncogene homolog (Kirsten Sıçan Sarkomu viral onkogen homologu)

* : Bağımsız İki Örneklem t-testi; SS: Standart Sapma

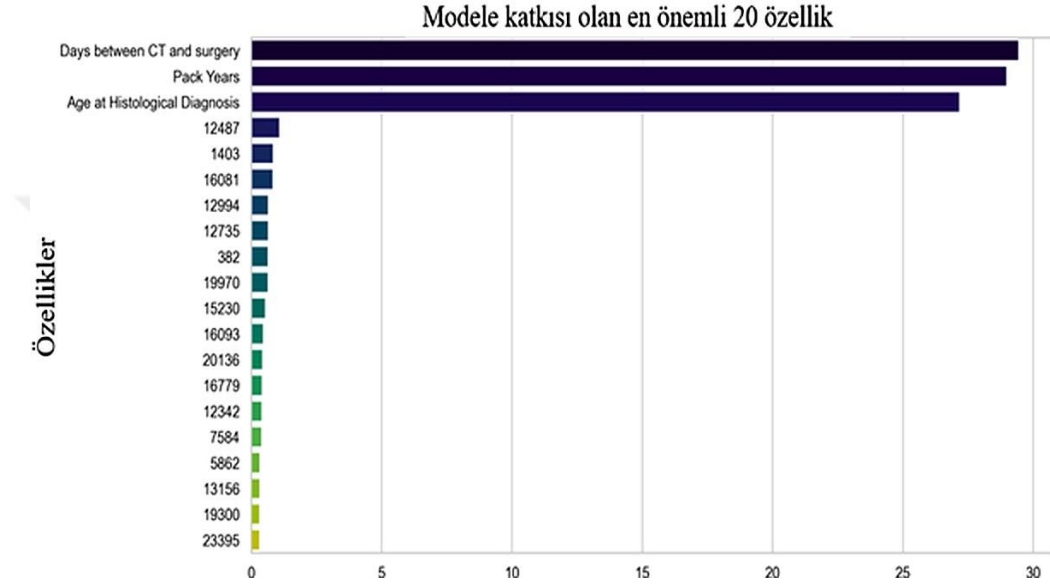
Veri setinde EGRF mutasyona sahip 43 hasta ve EGRF mutasyonu olmayan 168 örnek ele alındı. Ayrıca KRAS mutasyona sahip 38 hasta ve KRAS mutasyon olmayan 173 örnek hesaplandı. KHDAK hastalarının niceliksel biyobelirteçlerine ilişkin ortalama $\pm$ standart sapma ve her bir değişkene ait p-değerleri Tablo 4.2'de gösterildi.

EGFR ve KRAS mutasyon durumlarına göre Tablo 4.2 incelendiğinde, histolojik tanı yaşı her iki mutasyon durumuna göre (68 ± 10 , 68 ± 10 , $p=0.924$) ve ağırlık (122 ± 18 , 123 ± 22 , $p=0.783$) değişkenlerinde gruplar arasında istatistiksel olarak anlamlı bir fark bulunmadı ($p>0.05$). “Pack Years değişkeni, hastaların yaşam boyu sigara içme yükünü göstermekte olup, günlük tüketilen sigara paketi sayısı ile sigara içilen toplam yıl sayısının çarpımı olarak hesaplanmıştır (56).” Pack Years değişkeni açısından EGFR mutasyonu olmayan grubun ortalaması daha büyük (32 ± 16 , 41 ± 24) ve bu fark istatistiksel olarak anlamlı bulundu ($p<0.05$; $p=0.004$). BT ve ameliyat arasındaki günler açısından ise gruplar arasında anlamlı bir fark bulunmadı (39 ± 27 , 53 ± 68 , $p=0.209$). KRAS mutasyon durumu değerlendirildiğinde histolojik tanı yaşı (66 ± 10 vs 68 ± 10 , $p=0.165$), hastanın ağırlığı (128 ± 26 , 121 ± 20 , $p=0.096$), Pack Years (43 ± 28 , 38 ± 21 , $p=0.257$) ve BT ile ameliyat arasındaki günler (48 ± 39 , 51 ± 67 , $p=0.789$) değişkenleri istatistiksel olarak anlamlı bulunmadı. KHDAK hastalarına ilişkin DICOM görüntülerinden elde edilen süper piksellere ayrılmış görüntüler, otsu eşikleme ve tümörün olduğu bölgenin sınırları kırmızı renkle gösterildi ve referans noktalar belirlendi.(Şekil 4.1)



Şekil 4.1. KHDAK hastalarına ilişkin DICOM görüntülerine ilişkin segmentasyon, maskeleme ve kontur analizi

KHDAK görüntü işleme aşamasında (a) Grid: süper piksel segmentasyonu için kullanılan geometrik bir ağ yapısı olarak tanımlanmıştır; (b) orijinal gri tonlamalı DICOM görüntüsünü ifade etmektedir; (c) tümör bölgesini beyaz renkte temsil eden ikili segmentasyon maskesi; ve (d) değerlendirme ve sınıflandırma verileriyle birlikte kırmızı renk ile sınırları belirleyerek nihai referans noktaların analizi temsil edildi. EGFR mutasyonuna ilişkin modele katkısı en fazla olan klinik özellikler ve her bir görüntünün farklı özellikleri verildi (Şekil 4.2).



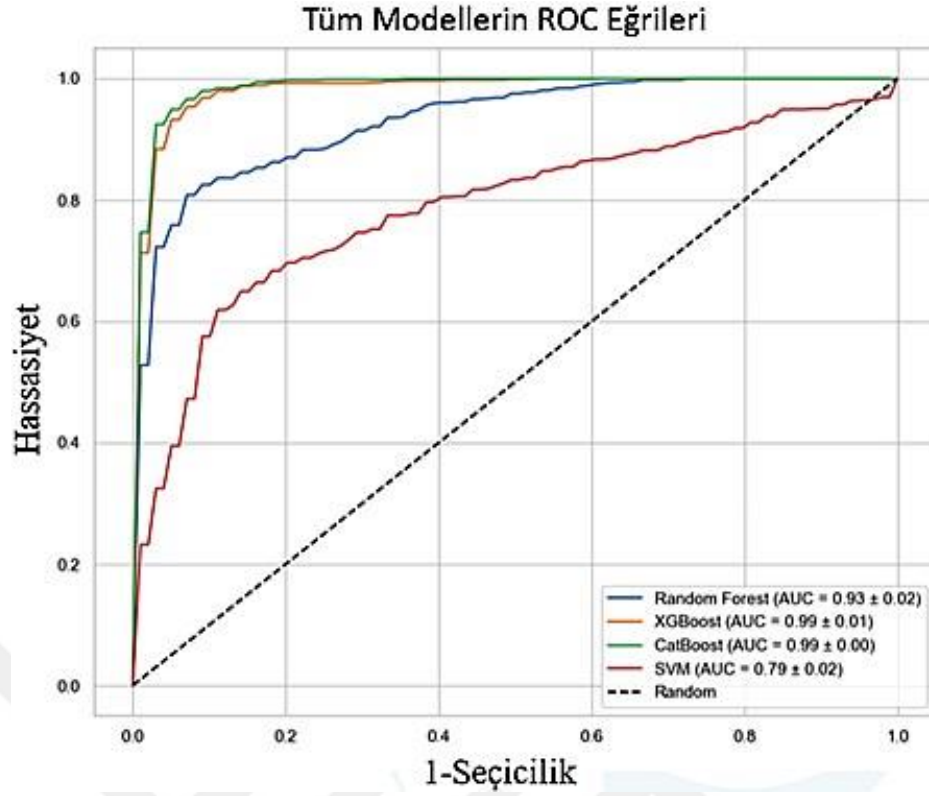
Şekil 4.2. EGFR mutasyonuna ilişkin en önemli değişkenlerin belirlenmesi

Şekil 4.2’de, ML modellerine en fazla katkı sağlayan ilk 20 değişken sunuldu. En önemli üç değişken olarak CT taraması ile cerrahi arasındaki gün sayısı, hastaların yaşam boyu sigara içme yükünü göstermekte olup, günlük tüketilen sigara paketi sayısı ile sigara içilen toplam yıl sayısının çarpımı olarak hesaplanan “Pack Years” ve histolojik tanı yaşı olarak hesaplandı. İlk üç değişkenin katkısı diğer değişkenlere kıyasla belirgin şekilde daha yüksek olup, CNN mimarisi olarak 16 katmandan oluşan VGG16 DÖ yöntemi ile son katmanda elde edilen nicel değişkenler, nispeten düşük etkiye sahip olduğu görüldü. Modelin karar sürecinde modele katkısı en çok olan üç değişkenin etkili olduğu ve tahmin gücünü büyük ölçüde belirlediği görüldü.

Tablo 4.3. EGFR mutasyon durumuna göre sınıflandırma performansına ilişkin metrikler

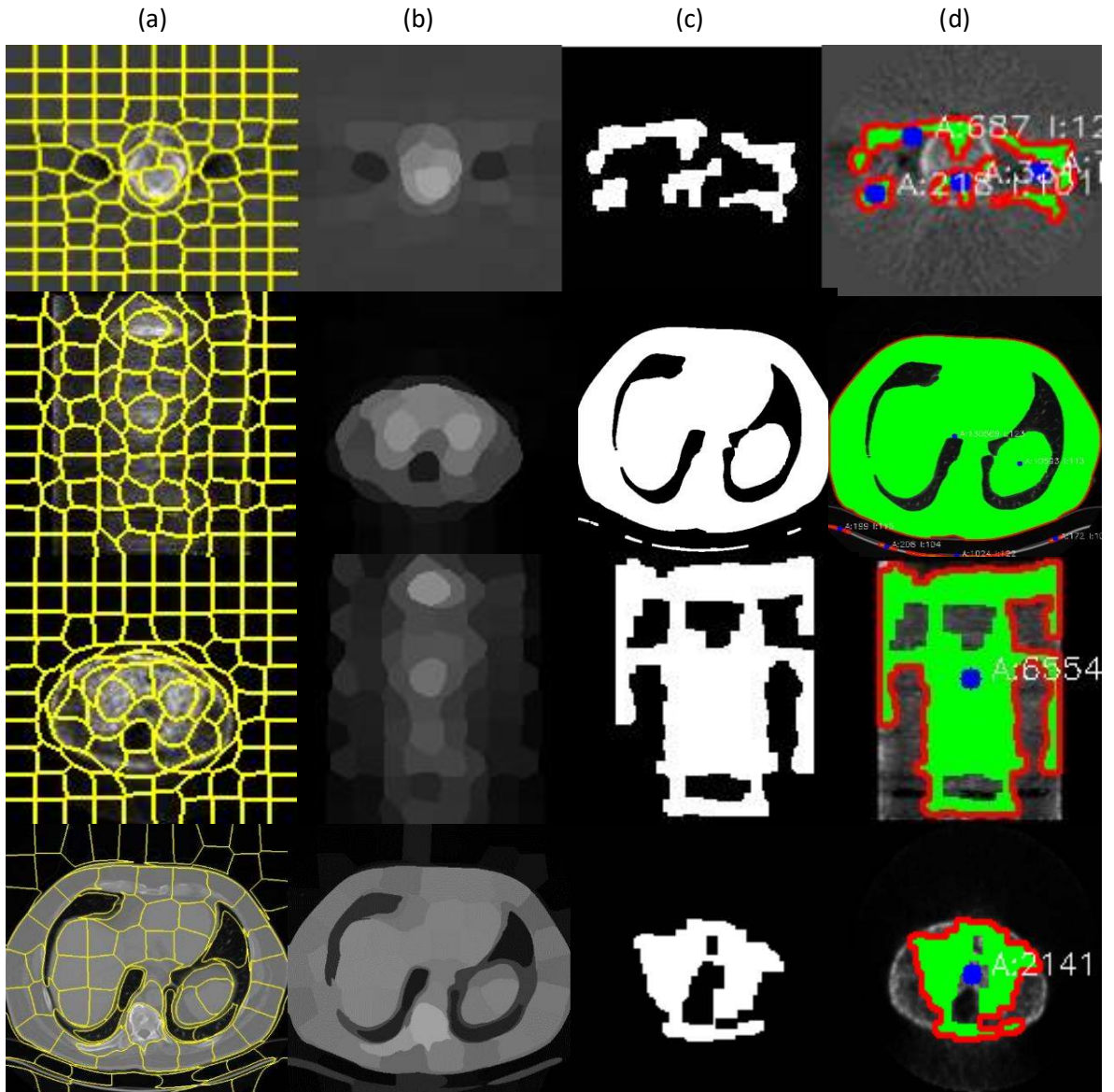
Model	Doğruluk	Dengelenmiş Doğruluk	Kesinlik	Duyarlılık	F1-Skor	ROC-AUC
CatBoost	0.967 ± 0.010 (0.966 - 0.967)	0.931 ± 0.022 (0.931 - 0.932)	0.968 ± 0.010 (0.968 - 0.969)	0.991 ± 0.003 (0.990 - 0.991)	0.979 ± 0.006 (0.979 - 0.980)	0.989 ± 0.002 (0.988 - 0.989)
Random Forest	0.888 ± 0.017 (0.887 - 0.888)	0.786 ± 0.038 (0.785 - 0.788)	0.907 ± 0.017 (0.907 - 0.908)	0.957 ± 0.006 (0.956 - 0.957)	0.931 ± 0.010 (0.931 - 0.932)	0.933 ± 0.021 (0.932 - 0.934)
SVM	0.730 ± 0.015 (0.730 - 0.731)	0.714 ± 0.015 (0.713 - 0.714)	0.903 ± 0.011 (0.902 - 0.903)	0.742 ± 0.028 (0.741 - 0.743)	0.814 ± 0.014 (0.813 - 0.815)	0.786 ± 0.017 (0.786 - 0.787)
XGBoost	0.963 ± 0.010 (0.963 - 0.964)	0.931 ± 0.012 (0.930 - 0.931)	0.969 ± 0.004 (0.968 - 0.969)	0.985 ± 0.009 (0.985 - 0.986)	0.977 ± 0.006 (0.976 - 0.977)	0.985 ± 0.007 (0.984 - 0.985)

DICOM görüntülerinden elde edilen nicel veriler ve KHDAK hastalarına ilişkin klinik verilerin entegre edilmesi ile geliştirilen sınıflandırma modelleri ve her bir metrik için 5-katlı çapraz geçerlilik sonucunda elde edilen performans metrikleri ile 5 farklı doğruluk, dengelenmiş doğruluk, kesinlik, duyarlılık, F1-Skor ve ROC-AUC metrikleri (ortalama $\pm$ standart sapma) Tablo 4.3'te gösterildi. XGBoost ML modeli 0.967 ± 0.010 doğruluk değeri ve (0.966 - 0.967) %95 güven aralığı ile en az standart sapma ile en iyi sınıflandırma performansına sahip olduğu görüldü. Modellerin performansı; doğruluk, dengelenmiş doğruluk, kesinlik, duyarlılık, F1-Skor ve ROC-AUC metrikleri ile değerlendirildi. ROC eğrisine ilişkin EGFR mutasyon bilgisi aşağıda verildi (Şekil 4.3).



Şekil 4.3. ROC eğrilerini kullanarak makine öğrenmesi model performanslarının karşılaştırılması

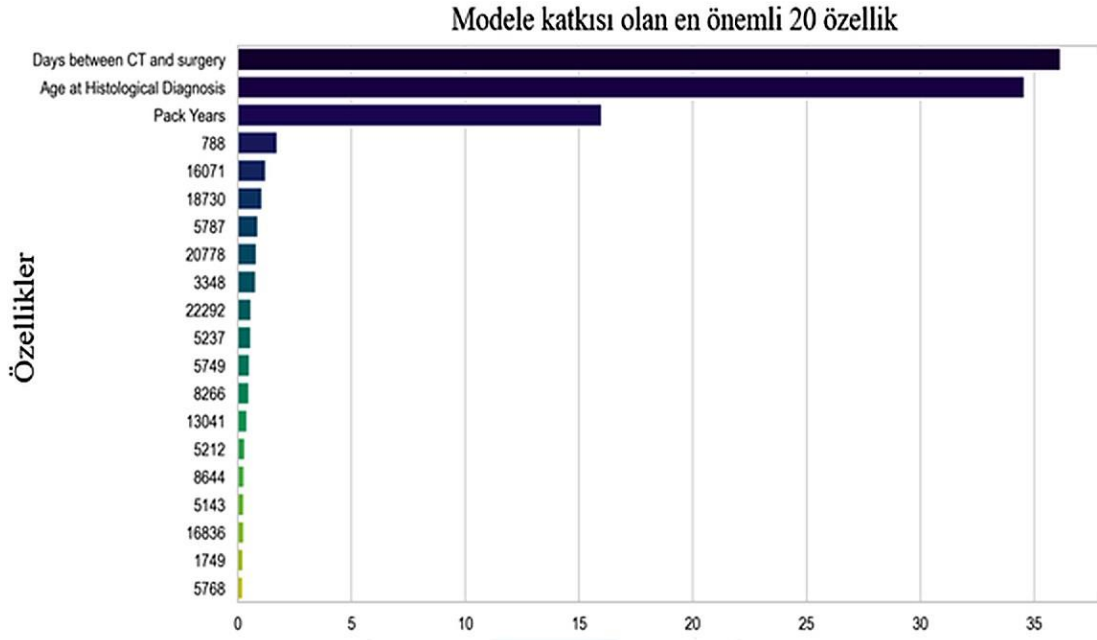
Şekil 4.3'te iki sınıflı EGRF mutasyon durumu hedef değişkeninin, ML modellerini karşılaştırmak için 5-katlı çapraz geçerlilik sonucunda elde edilen ortalama ve standart sapmaya sahip AUC değerleri hesaplandı. XGBoost ve CatBoost algoritmalarının sırasıyla AUC değerleri (0.99 ± 0.01) ve (0.99 ± 0.00) olarak hesaplandı. Ayrıca minimum standart sapma ile modele ilişkin tahminlerin kararlı ve güvenilir olduğunu gösterildi. KHDAK hastalarına ilişkin DICOM görüntülerinden elde edilen bazı çıktılar aşağıda gösterilmiştir (Şekil 4.4).



Şekil 4.4. Akciğer görüntülerinde segmentasyon, maskeleme ve kontur analizi

Görüntülerden elde edilen tümör bölgeleri yeşil renk ile belirlendi. Tümör merkezleri mavi renk ile işaretlendi ve ayrıca tümör sınırlarını belirlemek için kontur kullanıldı. Konturlar kırmızı çizgilerle belirlendi. KHDAK görüntü işleme aşamasında (a) Grid: Süper piksel segmentasyonu için kullanılan geometrik bir ağ yapısını ifade eder; (b) orijinal gri tonlamalı DICOM görüntüsü gösterildi; (c) tümör bölgesini beyaz renkte temsil eden ikili segmentasyon maskesi; ve (d) değerlendirme ve sınıflandırma verileriyle birlikte renkli bir görüntü içeren nihai analiz sonucu verildi.

KRAS mutasyonuna ilişkin modele katkısı en fazla olan klinik faktörler ve her bir görüntünün farklı özellikleri verildi (Şekil 4.5).



Şekil 4.5. KRAS mutasyonuna ilişkin en önemli değişkenlerin belirlenesi

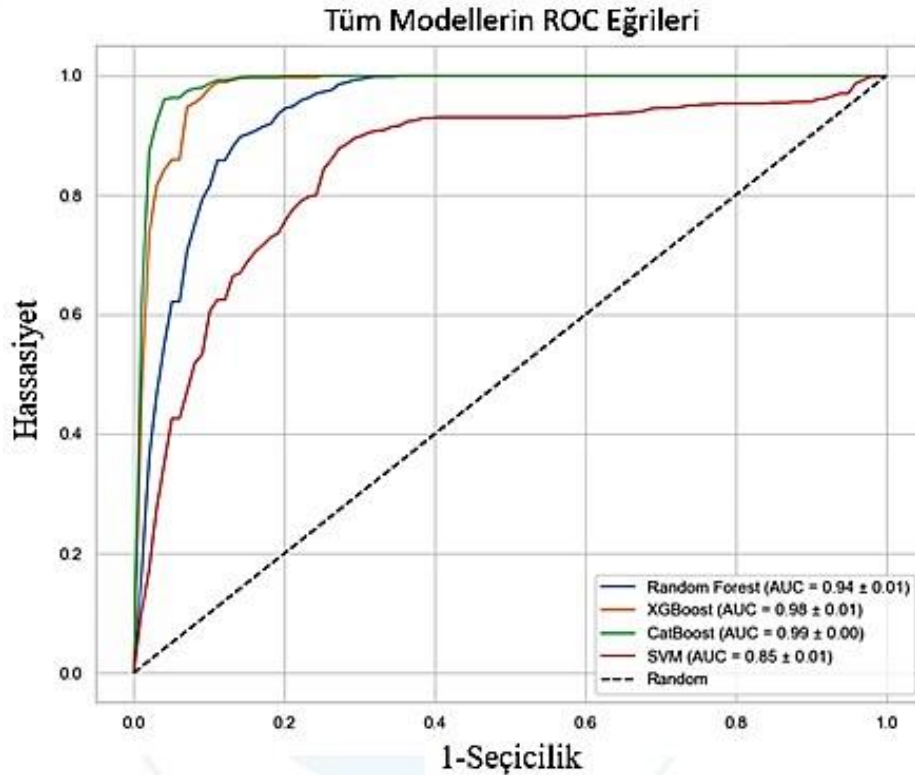
Modele en önemli katkıyı sağlayan değişkenler sırasıyla CT ile ameliyat arasında geçen gün sayısı ve histolojik tanı yaşı olarak hesaplandı. Bu faktörleri, paket yılı (sigara kullanım süresi) takip etmektedir. DICOM görüntülerinden elde edilen CNN mimarisi olarak 16 katmandan oluşan VGG16 derin öğrenme yöntemi ile elde edilen nicel değişkenler (16071, 18730, 5787, 20778, 3348, 22292, 5237, 5740, 8266, 13041, 5212, 8644, 5143, 16836, 1749, 5768) ise modele klinik verilerden daha düşük katkı sağlamakla birlikte, bu nicel veriler son katmanda elde edilen görüntülerin farklı özelliklerini temsil edildiği gösterildi. Şekil 4.5'te gösterilen bu 20 değişken arasında, klinik değişkenlerin (ilk üç değişken) modele katkısı, görüntü tabanlı değişkenlerden daha yüksek olduğu görüldü.

Tablo 4.4. KRAS mutasyon durumuna göre sınıflandırma performansına ilişkin metrikler

Model	Doğruluk	Dengelenmiş Doğruluk	Kesinlik	Duyarlılık	F1-Skor	ROC-AUC
CatBoost	0.965 ± 0.015 (0.964 - 0.966)	0.954 ± 0.021 (0.953 - 0.955)	0.953 ± 0.024 (0.952 - 0.954)	0.994 ± 0.007 (0.994 - 0.995)	0.973 ± 0.011 (0.973 - 0.974)	0.990 ± 0.005 (0.990 - 0.991)

Random Forest	0.879 ± 0.008 (0.878 - 0.879)	0.839 ± 0.010 (0.839 - .840)	0.848 ± 0.012 (0.848 - 0.849)	0.986 ± 0.020 (0.985 - 0.987)	0.912 ± 0.007 (0.911 - 0.912)	0.942 ± 0.008 (0.942 - 0.942)
SVM	0.813 ± 0.024 (0.812 - 0.814)	0.795 ± 0.018 (0.794 - 0.796)	0.846 ± 0.010 (0.845 - 0.846)	0.863 ± 0.043 (0.861 - 0.864)	0.854 ± 0.022 (0.853 - 0.855)	0.847 ± 0.013 (0.847 - 0.848)
XGBoost	0.951 ± 0.006 (0.950 - 0.951)	0.936 ± 0.007 (0.935 - 0.936)	0.935 ± 0.009 (0.934 - 0.935)	0.991 ± 0.011 (0.991 - 0.992)	0.962 ± 0.005 (0.962 - 0.962)	0.982 ± 0.011 (0.981 - 0.982)

DICOM görüntülerinden elde edilen nicel veriler ve KHDAK hastalarına ilişkin klinik verilerin entegre edilerek 5-kat çapraz doğrulama sonucunda performans metrikleri (ortalama ± standart sapma) Tablo 4.4'te olduğu gibi hesaplandı. CatBoost ML modeli (0.965 ± 0.015) doğruluk değeri ve standart sapma ($0.964 - 0.966$) %95 güven düzeyi ile en iyi sınıflandırma performansına sahip olduğu görüldü. Modellerin performansı; doğruluk, dengelenmiş doğruluk, kesinlik, duyarlılık, F1-Skor ve ROC-AUC metrikleri ile değerlendirildi. KRAS mutasyon durumuna ilişkin ML modellerini karşılaştırmak için modellerin ortalama AUC ve standart sapmaları hesaplandı (Şekil 4.6).



Şekil 4.6. ROC eğrilerini kullanarak makine öğrenimi model performanslarının karşılaştırılması

CatBoost ve XGBoost modellerine ilişkin AUC değerleri sırasıyla 0.980 ± 0.01 ve 0.99 ± 0.00 olarak hesaplandı. Minimum standart sapma ile modele ilişkin tahminlerin kararlı ve güvenilir olduğu gösterildi. CatBoost ve XGBoost modelleri düşük yanlış pozitif oranı ve yüksek doğru pozitif oranı ile en başarılı modeller olarak hesaplandı. EGFR ve KRAS mutasyon durumlarına ilişkin klinik veriler, ML ortalama sınıflandırma performans metrikleri ve DICOM görüntü verileri ile klinik verilerin entegrasyonu ile elde edilen ML sınıflandırmasına ilişkin tahmin edilen performans metrikleri ve bu değişimin oranları Tablo 4.5'te sunuldu.

Tablo 4.5. EGFR ve KRAS mutasyonlarının sınıflandırılmasına ilişkin tablo

Mutasyon	Model	Metrik	Klinik Veri	Klinik Veri + Görüntü	Değişim
EGFR	CatBoost	ROC-AUC	0.890	0.989	%11.1
		Doğruluk	0.823	0.967	%17.5
		Dengelenmiş Doğruluk	0.814	0.931	%14.4
		Kesinlik	0.825	0.968	%17.3
		Duyarlılık	0.867	0.991	%14.3
			0.845	0.979	%15.9
			0.890	0.985	%10.7
			0.823	0.963	%17.0
			0.814	0.931	%14.4
		F1-Skor	0.825	0.969	%17.5
	XGBoost	ROC-AUC	0.867	0.985	%13.6
		Doğruluk	0.845	0.977	%15.6
		Dengelenmiş Doğruluk	0.890	0.933	%4.8
			0.823	0.888	%7.9
		Kesinlik	0.814	0.786	% -3.4
		Duyarlılık	0.825	0.907	%9.9
		F1-Skor	0.867	0.957	%10.4
		Doğruluk	0.845	0.931	%10.2
	RF	Dengelenmiş Doğruluk			
		Kesinlik			

KRAS	SVM	ROC-AUC	0.890	0.786	% -11.7
		Doğruluk	0.823	0.730	% -11.3
				0.714	% -12.3
		Dengelenmiş Doğruluk	0.814	0.903	%9.5
		Kesinlik	0.825	0.742	% -14.4
		Duyarlılık	0.867	0.814	% -3.7
		F1-Skor	0.845	0.982	%20.9
	XGBoost	ROC-AUC	0.812	0.951	%20.5
		Doğruluk	0.789	0.936	%20.3
		Dengelenmiş Doğruluk	0.778	0.935	%22.2
		Kesinlik	0.765	0.991	%22.0
		Duyarlılık	0.812	0.962	%21.2
		F1-Skor	0.788	0.990	%21.9
		ROC-AUC	0.812	0.965	%22.3
	CatBoost	Doğruluk	0.789	0.954	%22.6
		Dengelenmiş Doğruluk	0.778	0.953	%24.6
		Kesinlik	0.765	0.994	%22.4
		Duyarlılık	0.812	0.973	%23.5
		F1-Skor	0.788	0.942	%16.0
		ROC-AUC	0.812	0.879	%11.4
		Doğruluk	0.789	0.839	%7.8
	RF			0.848	%10.8
		Dengelenmiş Doğruluk	0.778	0.986	%21.4
		Kesinlik	0.765	0.912	%15.7
		Duyarlılık	0.812	0.847	%4.3
		F1-Skor	0.788	0.813	%3.0
		ROC-AUC	0.812	0.795	%2.2
		Doğruluk	0.789	0.846	%10.6
SVM	SVM			0.863	%6.3
		Dengelenmiş Doğruluk	0.778	0.854	%8.4
		Kesinlik	0.765		
		Duyarlılık	0.812		
		F1-Skor	0.788		

5. TARTIŞMA

KHDAK'ın yapısal özelliği, Son zamanlarda uzmanların hasta bireylerin tedavi sürecinde ve karar alma sürecinde avantaj sağlamaktadır. EGFR ve KRAS mutasyonlarının sağlık uzmanlarının tedaviye yönelik hedeflerini olumlu yönde geliştirmesine ve kişiselleştirilmiş tedavide uygulamasına olanak sağlamaktadır. Ancak tekrarlayan biyolojiklerin invaziv olması ve tümör heterojenitesi, genomik profillemenin klinikte karşılaşılan önemli sorunlardır. Radyomik, yüksek verimli radyografik görüntülerden birçok görüntü özelliği çıkarma, biyobelirteçleri tahmin etme ve invazif olmayacak bir şekilde lezyonların fenotipini tahmin etmek için pek çok özellik uygular. Radyomik özelliklerin, tümör fenotiplerini seçme ve biyobelirteçleri tahmin etme gibi akciğer kanserinin önlenmesinde kullanılabildiğine ilişkin çok sayıda araştırma yapılmıştır. Ultrason ve MRI gibi invazif olmayan görüntüleme teknikleri ile kanser gibi hastalıkların teşhisinde kullanılır (86).

Görüntüleme teknikleri ile sağlık uzmanları, hastalığın türünü belirlemede ve hastalığın yayılım evrelerini kestirip erken tanı ve tedavi sürecinde yol gösterici olarak kullanılmaktadır. İnvazif olmayan yöntemler (CT, MRI, PET) gibi görüntü yöntemleri tümör boyutunu, bir görüntüdeki tümör sayısını ve yoğunluğunu belirlemek için kullanılır. Görüntüleme yöntemleri sadece akciğer hastalığının varlığını tespit etmekle kalmaz, aynı zamanda prognostik (Hastalığın seyri, Tedavi yanıtı, Yaşam beklentisi) gibi biyolojik veya klinik biyobelirteçleri hakkında önemli bilgiler sunmaktadır (87).

Bu tez çalışmasında elde edilen veriler, KHDAK'lı hastalarda klinik verilere ilişkin EGFR ve KRAS mutasyonlarının durumu, DICOM görüntülerinde tümörün tespiti için çeşitli ML yaklaşımı ile bu mutasyonların erken tespiti amacıyla model performanslarını değerlendirmektir. Kişiselleştirilmiş tıp ve hedefe yönelik tedavi seçeneklerinin uygulanmasına yardımcı olabilir. KHDAK'lı 211 hastadan oluşan genel klinik veri seti ve biyomedikal görüntü veri setlerinden oluşan The Cancer Imaging Archive (TCIA) üzerinde yapılan çalışmadan elde edilmiştir (88, 89). Literatürde aynı veri setini kullanarak 211 KHDAK'lı hastadan 161 kohort hastası eklenek yapılan bir çalışmada EGRF ve KRAS mutasyonlarının tahmini için 10 kat çapraz doğrulama ile XGBoost modeline ilişkin EGRF ve KRAS sırasıyla 0.83 ve 0.86 olarak elde edilmiş ayrıca mutasyonlara ilişkin AUC-ROC değerleri sırasıyla 0.89 ve 0.812 elde edilmiştir (90). KHDAK'da EGFR ve KRAS mutasyonlarının tahmini için ML tabanlı yapılan

çalışma ile referans çalışmanın AUC değerleri karşılaştırıldığında, çalışmamızda 2.231 tıbbi görüntüden elde edilen radyomik değişkenler ile klinik verilerin entegrasyonu sayesinde KHDAK hastalarında hem EGFR hem de KRAS mutasyonlarının tahmininde önemli ölçüde performans gösterildi. Bu çalışmanın EGFR mutasyon durumu için CatBoost (AUC = 0.99 ± 0.00) ve XGBoost (AUC = 0.99 ± 0.01), KRAS mutasyonları durumu ise CatBoost (AUC = 0.99 ± 0.00) ve XGBoost (AUC = 0.98 ± 0.01) kullanan yaklaşımımız, tahmin doğruluğunda belirgin bir iyileşme göstermektedir. Referans çalışmanın 10 kat çapraz doğrulama yaklaşımına kıyasla bu çalışmada model performanslarının doğru değerlendirilmesi için 5 katlı çapraz geçerlilik kullanıldı. Modellerin minimum standart sapma ile tutarlı ve güvenilir olduğu saptanmıştır. Elde edilen üstün performans, klinik değişkenlere ilişkin bilgiler ile kapsamlı radyomik değişkenlerle birleştirilmesine dayanarak, bütünleşik veri stratejisine ve aynı zamanda gelişmiş gradient boosting algoritmalarının uygulanmasına dayanmaktadır. Bu entegre yaklaşım, sadece ML modellerine ilişkin tahmin doğruluğunu artırmakla kalmayıp, aynı zamanda farklı veri bölünmelerinde güçlü genelleme yetenekleri göstermektedir.

ROC eğrileri analizi neticesinde EGFR ve KRAS mutasyon durumlarını CatBoost ve XGBoost modellerine ilişkin performans metrikleri düşük yanlış pozitif oranı ve yüksek doğru pozitif oranı ile tahmin etmede en başarılı modeller olduğunu ortaya koymaktadır. CNN mimarisinde 16 katmandan oluşan VGG16 derin öğrenme yöntemi ile, görüntülerden değişkenler çıkarılarak, KHDAK klinik veri seti ve görüntü verileri ile entegre edilmesi, eksik gözlemlerde kategorik değişkenler için “mod” ve nicel değişkenler ise “medyan” yöntemi ile atama yapıldı. RF algoritması ile değişken seçimi yapılmıştır. EGFR mutasyon durumuna ilişkin modelin performansı değerlendirildiğinde (Tablo 4.3), çalışmada mutasyon durumunu tahmin etmede dört ML modeli karşılaştırılmış ve en yüksek performansı CatBoost modeli göstermiştir.

CatBoost modelinin doğruluk (0.967 ± 0.010), dengelenmiş doğruluk (0.931 ± 0.022), kesinlik (0.968 ± 0.010), duyarlılık (0.991 ± 0.003), F1-Skoru (0.979 ± 0.006) ve ROC-AUC (0.989 ± 0.002), Random Forest modelinin doğruluk (0.888 ± 0.017), dengelenmiş doğruluk (0.786 ± 0.038), kesinlik (0.907 ± 0.017), duyarlılık (0.957 ± 0.006), F1-Skoru (0.931 ± 0.010) ve ROC-AUC (0.933 ± 0.021), SVM modelinin doğruluk (0.730 ± 0.015), dengelenmiş doğruluk (0.714 ± 0.015), kesinlik (0.903 ± 0.011), duyarlılık (0.742 ± 0.028), F1-Skoru (0.814 ± 0.014) ve ROC-AUC (0.786 ± 0.017), CatBoost modeli daha yüksek performans sağlamıştır. CatBoost modeline çok yakın performans değerleriyle XGBoost modeli takip etmiş olup, bu model de doğruluk

(0.963 $\pm$ 0.010), kesinlik (0.969 $\pm$ 0.004) ve ROC-AUC (0.985 $\pm$ 0.007) metrikleri açısından oldukça başarılı sınıflandırma performansı göstermektedir. CatBoost ve XGBoost modellerinin sırasıyla 0.991 ve 0.985 gibi yüksek duyarlılık değerler ölçülmüştür. Bu durumda, yanlış negatif değerlerinin minimize ettiğini ve tüm performans metriklerinin düşük standart sapma değerlerine sahip olması modellerin güvenilir ve tutarlı olduğunu göstermektedir.

KRAS mutasyon durumuna ilişkin modelin performansı incelendiğinde (Tablo 4.4), ML performans metriklerinden yüksek performansa sahip modelin CatBoost modelinde gözlenmektedir. CatBoost modeli, doğruluk, dengelenmiş doğruluk, kesinlik, duyarlılık, F1-skor ve ROC-AUC metrikleri sırasıyla; (0.965 $\pm$ 0.015), (0.954 $\pm$ 0.021), (0.953 $\pm$ 0.024), (0.994 $\pm$ 0.007), (0.973 $\pm$ 0.011) ve (0.990 $\pm$ 0.005) olarak hesaplanmıştır. Random Forest modeli doğruluk 0.879 $\pm$ 0.008 (0.878 - 0.879), dengelenmiş doğruluk 0.839 $\pm$ 0.010 (0.839 - .840), kesinlik 0.848 $\pm$ 0.012 (0.848 - 0.849), kesinlik 0.986 $\pm$ 0.020, (0.985 - 0.987), F1-skor 0.912 $\pm$ 0.007 (0.911 - 0.912) ve ROC-AUC 0.942 $\pm$ 0.008 (0.942 - 0.942) hesaplanmıştır. SVM modeli, performans metrikleri; doğruluk, dengelenmiş doğruluk, kesinlik, duyarlılık, F1-skor ve ROC-AUC metrikleri sırasıyla; 0.813 $\pm$ 0.024 (0.812 - 0.814), 0.795 $\pm$ 0.018 (0.794 - 0.796), 0.846 $\pm$ 0.010 (0.845 - 0.846), 0.863 $\pm$ 0.043 (0.861 - 0.864), 0.854 $\pm$ 0.022 (0.853 - 0.855), ve 0.847 $\pm$ 0.013 (0.847 - 0.848) şeklinde hesaplanmıştır. XGBoost modeline ilişkin doğruluk değeri 0.951 $\pm$ 0.006 (0.950 - 0.951), kesinlik 0.935 $\pm$ 0.009 (0.934 - 0.935) ve ROC-AUC 0.982 $\pm$ 0.011 (0.981 - 0.982) değerleri CatBoost algoritmasına yakın performans sergilemektedir. İlgili tabloda CatBoost ve XGBoost modellerinin duyarlılık değerleri sırasıyla incelendiğinde (0.994 ve 0.991) “modellerin yanlış negatif sonuçları minimuma indirdiği” yorumu yapılabilir. İncelenen metriklerin tamamında standart sapmanın düşük olduğu gözlenmiş ve bu da modelin tutarlı ve güvenilir olduğunu göstermektedir.

Klinik verilere ilişkin performans metrikleri ve görüntü verileri ile entegre edilen sonuçlar (Tablo 4.5) karşılaştırıldığından hem EGFR hem de KRAS mutasyon durumlarını kestirmede görüntü veri setine ilişkin değişkenlerin birleştirilmesi model performansında önemli iyileştirmeler sağlandığı gözlemlendi. EGFR mutasyon durumu için, CatBoost ve XGBoost modellerinin performans değerleri benzer olarak elde edildi. CatBoost ve XGBoost modeline ilişkin ROC-AUC performans değerleri ve iyileştirme oranları sırasıyla; 0.812'den 0.990'a (+21.9%) ve 0.812'den 0.982'ye (+20.9%) yükseldi. CatBoost ve XGBoost model doğruluk performans değerleri ve iyileştirme oranları

sırasıyla; 0.789'dan 0.951'e (20.5%) ve 0.789'dan 0.965'e (22.3%) yükseltildi. EGFR mutasyonuna ilişkin iyileştirme sonucunda, RF modeli orta düzeyde ve SVM modeli ise negatif performansa sahiptir. Tablo 4.5'te elde edilen sonuçlar CatBoost ve XGBoost gradient boosting modelleri, EGFR ve KRAS mutasyonlarında tutarlı ve yüksek performansa sahip olduğu görüldü.



6. SONUÇ VE ÖNERİLER

Bu tez çalışmasında, KHDAK hastaların biyomedikal görüntüler ve ilgili görüntülere ilişkin klinik veriler kullanılarak çeşitli makine öğrenmesi algoritmaları ile modellerin performans metrikleri incelendi. Model performans metrikleri, sadece klinik veri, biyomedikal veriler ile görüntü ilişkin değişkenlerin birleştirilmesi ile elde edilen sonuçlar ve biyomedikal görüntü ve klinik değişkenlerin birleşmesiyle sağlanan iyileştirme oranları karşılaştırıldı. Modellerin sınıflandırma performansları doğruluk, dengelenmiş doğruluk kesinlik, duyarlılık, F-skoru, ROC-AUC metrikleri ile değerlendirildi.

Bulgular incelendiğinde, en düşük sınıflandırma performansı destek vektör makineleri modelinden elde edildi. Diğer taraftan biyomedikal klinik ve görüntü veri setinde EGFR mutasyon durumunu tahmin etmede CatBoost, algoritması ve KRAS mutasyon durumunu tahmin etmede de CatBoost modeli en iyi sınıflandırma performansı sağlaması ile beraber KRAS mutasyon durumunun tahmin edilmesinde XGBoost modeli de başarılı sınıflandırma performansı gösterdi.

Biyomedikal görüntü ve klinik verilerinde yapılacak olan çalışmalarda, gradient boosting modellerine dayalı CatBoost ve XGBoost kullanılması, görüntü verileri ile klinik verilerin entegrasyonunun standart hale getirilmesi, görüntülerden farklı değişken seçme yöntemleri kullanılarak modellere ilişkin tahmin performans değerlerinin artırılması, makine öğrenme modellerinde metriklerin ortalama ve standart sapma değerleri belirlenerek modelin tutarlı, kararlı ve güvenliğinin belirlenmesi önerilmektedir.

KAYNAKLAR

1. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RG, Granton P, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441-46.
2. Griethuysen JJM, Andriy F, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):104-107.
3. Chaddad A, Daniel P, Sabri S, Desrosiers C, Abdulkarim B. Integration of radiomic and multi-omic analyses predicts survival of newly diagnosed IDH1 wild-type glioblastoma. *Cancers (Basel)*. 2019;11(8):1148.
4. Ren Y, Zang X, Rui W, Pang H, Qiu T, Wang J, et al. Noninvasive prediction of IDH1 mutation and ATRX expression loss in low- grade gliomas using multiparametric MR radiomic features. *J Magn Reson Imaging*. 2019;49(3):808-17.
5. Sakai Y, Yang C, Kihira S, Tsankova N, Khan F, Hormigo A, et al. MRI radiomic features to predict IDH1 mutation status in gliomas: A machine learning approach using gradient tree boosting. *Int J Mol Sci*. 2020;21(21):8004:2-22
6. Yogananda CGB, Shah BR, Yu FF, Pinho MC, Nalawade SS, Murugesan GK, et al. A novel fully automated MRI-based deep-learning method for classification of 1p/19q co-deletion status in brain gliomas. *Neurooncol Adv*. 2020;2(Supplement_4):iv42-iv8.
7. Choi YS, Bae S, Cang JH, Kang SG, Kim SH, Kim J, et al. Fully automated hybrid approach to predict the IDH mutation status of gliomas via deep learning and radiomics. *Neuro Oncol*. 2021;23(2):304-13.
8. Li Y, Liu X, Qian Z, Sun Z, Xu K, Wang K , et al. Genotype prediction of ATRX mutation in lower-grade gliomas using an MRI radiomics signature. *Eur Radiol*. 2018;28:2960-968.
9. Haubold J, Demircioglu A, Gratz M, Glas M, Wrede K, Sure U, et al. Non-invasive tumor decoding and phenotyping of cerebral gliomas utilizing multiparametric 18 F-FET PET-MRI and MR fingerprinting. *Eur J Nucl Med Mol Imaging*. 2020;47:1435-445.

10. Song L, Zhu Z, Mao L, Li X, Han W, Du H, et al. Clinical, conventional CT and radiomic feature-based machine learning models for predicting ALK rearrangement status in lung adenocarcinoma patients. *Front Oncol.* 2020;10:369.
11. Xu Y, Hosny A, Zeleznik R, Parmar C, Coroller T, Franco I, et al. Deep learning predicts lung cancer treatment response from serial medical imaging. *Clin Cancer Res.* 2019;25(11):3266-275.
12. Yamamoto S, Kom RL, Oklu R, Migdal C, Gotway MB, Weiss GJ, et al. ALK molecular phenotype in non-small cell lung cancer: CT radiogenomic characterization. *Radiology.* 2014;272(2):568-76
13. Shi L, He Y, Yuan Z, Benedic S, Valicenti R, Qiu J, et al. Radiomics for response and outcome assessment for non-small cell lung cancer. *Technol Cancer Res Treat.* 2018;17:1-14
14. Zhou M, Leung A, Echegaray S, Gentles A, Shrager JB, Jensen KC, et al. Non-small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications. *Radiology.* 2018;286(1):307-15.
15. Singh A, Chitalia R, Kontos D. Radiogenomics in brain, breast, and lung cancer: opportunities and challenges. *J Med Imaging.* 2021;8(3):1-14
16. Armanious K, Jiand C, Fischer N, Küstner T, Hepp T, Nikolaou K, et al. MedGAN: Medical image translation using GANs. *Comput Med Imaging Graph.* 2020;79:101684.
17. Razzak MI, Naz S, Zaib A. Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps: Automation of Decision Making.* 2018:323-50.
18. Yu D, Zhou M, Yang F, Dong D, Gevaert O, Liu Z. et al. Convolutional neural networks for predicting molecular profiles of non-small cell lung cancer. *IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017).* Piscataway (NJ): IEEE; 2017. p. 569–72.
19. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Trans Med Imaging.* 2016;35(5):1207-216.
20. Korfiatis P, Erickson B. Deep learning can see the unseeable: predicting molecular markers from MRI of brain gliomas. *Clin Radiol.* 2019;74(5):367-73.

21. Michael M. C, Mark D. H, Geoffrey M. The molecular biology of lung cancer. *Med J Aust.* 1991;154(8):501-03.
22. Jacobs C, van Rikxoort EM, Murphy K, Prokop M, Schaefer-Prokop CM, van Ginneken B. Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database. *Eur Radiol.* 2016;26:2139-147.
23. Chia MM, Holmes MD, McLennan G. The molecular biology of lung cancer. *Med J Aust.* 1991;154(8):501-03.
24. Paul A. B, Epidemiologic aspects of lung cancer. *CA Cancer J Clin.* 2003; 53:7-33
25. Siegfried, JM. Biology and chemoprevention of lung cancer. *Chest.* 1998;113(1):40-45.
26. Wei Q, Spitz MR. The role of DNA repair capacity in susceptibility to lung cancer: a review. *Cancer Metastasis Rev.* 1997;16:295-307.
27. Çakar E, Türker AY, Güleriyüz E, Karaca A. Detection of candidate nodules in lung tomography by image processing techniques. 21st National Biomedical Engineering Meeting IEEE.2017; 1-4
28. Munir K, Elahi H, Ayub A, Frezza F, Rizzi A. Cancer diagnosis using deep learning: A bibliographic review. *Cancers.* 2019;11(9):1235.
29. Sevim G, Üncü YA, Mercan T, Canpolat M. Image reconstruction for diffuse optical tomography using bi- conjugate gradient and transpose- free quasi minimal residual algorithms and comparison of them. *Int J Imaging Syst Technol.* 2021;31(4):1894-905.
30. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018;68(6):394-424.
31. Altarawneh M. Lung cancer detection using image processing techniques. *Leonardo Electron J Pract Technol.* 2012;11(21):147-58.
32. Altındış S, Morkoç İK. Sağlık hizmetlerinde büyük veri. *Ömer Halisdemir Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi.* 2018;11(2):257-71.
33. Milne R, Morly KI, Alamarri MA, Anwer S, Atutornu J, Baranova EE, et al. Demonstrating trustworthiness when collecting and sharing genomic data: Public views across 22 countries. *Genome Med.* 2021;13(1):92,1-12
34. Del Boccio P, Andrea Urbanı A. Homo sapiens proteomics: Clinical perspectives. *Ann Ist Super Sanita.* 2005;41(4):479-82.

35. Başaran E, Aras S, Cansaran-Duman D. General outlook and applications of genomics, proteomics and metabolomics. *Turk Hij Den Biyol.* 2010;67(2):85-96.
36. Karakaya H. Genetik, Samsun, Türkiye Klinikleri J Allergy-Asthma. Samsun, 2002;4:19-23.
37. Saxena S, Jena B, Gupta N, Das S, Sarmah D, Bhattacharya P, et al. Role of artificial intelligence in radiogenomics for cancers in the era of precision medicine. *Cancers.* 2022;14(12):2860.
38. Thawani R, McLane M, Beig N, Ghose S, Prasanna P, Velcheti V, et al. Radiomics and radiogenomics in lung cancer: a review for the clinician. *Lung cancer.* 2018;115:34-41.
39. Brahmer JR, Govindan R, Anders RA, Antonia SJ, Sagorsky S, Davies MJ, et al. The Society for Immunotherapy of Cancer consensus statement on immunotherapy for the treatment of non-small cell lung cancer (NSCLC). *J Immunother Cancer.* 2018;6(75):1-15.
40. Reck M, Rodríguez–Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, et al. Updated analysis of keynote-024: Pembrolizumab versus platinum-based chemotherapy for advanced non–small-cell lung cancer with PD-L1 tumor proportion score of 50% or greater. *J Clin Oncol.* 2019;37(7):537-46.
41. Johung KL, Yeh N, Desai NB, Williams TM, Lautenschlaeger T, Arvold ND, et al. Extended survival and prognostic factors for patients with ALK-rearranged non–small-cell lung cancer and brain metastasis. *J Clin Oncol.* 2016;34(2):123-29.
42. McAndrew A. An introduction to digital image processing with MATLAB: Course Technology Press, Boston, MA, United States; 2004,1-528
43. Samuel GA, Mayer CR, Michael MG, Geoffrey G, Anthony PR, Croft BY, et al. The reference image database to evaluate response to therapy in lung cancer (RIDER) project: A resource for the development of change- analysis software. *Clin Pharmacol Ther.* 2008;84(4):448-56.
44. Armato SG, Li F, Giger ML, MacMahon H, Sone S, Doi K. Lung cancer: performance of automated lung nodule detection applied to cancers missed in a CT screening program. *Radiology.* 2002;225(3):685-92.
45. Armato III SG, Giger ML, MacMahon H. Automated detection of lung nodules in CT scans: preliminary results. *Med Phys.* 2001;28(8):1552-561.
46. Wang C, Elazab A, Wu J, Hu Q. Lung nodule classification using deep feature fusion in chest radiography. *Comput Med Imaging Graph.* 2017;57:10-18.

47. Hornak JP. Teaching NMR using online textbooks. *Molecules*. 1999;4(12):353-65.
48. Nabiye VV. Yapay Zeka, Seçkin Yayıncılık San. Ve Tic AŞ, Ankara, 2003,1-724.
49. Bakator M, Radosav D. Deep learning and medical diagnosis: A review of literature. *Multimodal Technol Interact*. 2018;2(3):47.
50. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *Plos Med*. 2018;15(11):1-17.
51. Yasaka K, Abe O. Deep learning and artificial intelligence in radiology: Current applications and future directions. *Plos Med*. 2018;15(11):1-4.
52. Sunarti S, Rahman FF, Naufal M, Risky M, Febriyanto K, Masnina R. Artificial intelligence in healthcare: Opportunities and risk for future. *Gac Sanit*. 2021;35(1):67-70.
53. Russell SJ, Norvig P. Artificial intelligence: a modern approach: 3th ed. Malaysia, Pearson Education Limited; 2016 1-1152.
54. Karakaya A. Meme kanseri tahmininde makine öğrenmesi algoritmaları ve AutoML [yüksek lisans tezi]. Denizli: Pamukkale University; 2024.
55. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-44.
56. Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, et al. Data for NSCLC radiogenomics collection. *The Cancer Imaging Arc*.. 2017;K9
57. Fan C, Chen M, Wang X, Wang J, Huang B. A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Front Energy Res*. 2021;9:1-17.
58. Chu X, Ilyas IF, Krishnan S, Wang J, editors. Data cleaning: Overview and emerging challenges. *Proceedings of the 2016 International Conference on Management of Data*; 2016: 2201-206.
59. Singh D, Singh B. Investigating the impact of data normalization on classification performance. *Appl Soft Comput*. 2020;97:1-23.
60. Goble C, Stevens R. State of the nation in data integration for bioinformatics. *J Biomed Inform*. 2008;41(5):687-93.
61. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*. 2003;3(3):1157-182.
62. Aşlıyan R. Otsu ve Rocchio Metotlarıyla Beyin Tümörü Tespiti. *J Sci Technol* 2022; 43(10):69-74.

63. Huerta EB, Duval B, Hao J-K. A hybrid LDA and genetic algorithm for gene selection and classification of microarray data. *Neurocomputing*. 2010;73(13-15):2375-383.
64. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. *FEMS Microbiol Rev*. 2011;35(2):343-59.
65. Zebari R, Abdulazeez A, Zeebaree D, Zebari D, Saeed J. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *J Appl Sci Technol Trends*. 2020;1(1):56-70.
66. Emeç M, Özcanhan MH. Veri Ön İşleme ve Öznitelik Mühendisliğinin Yapay Zekâ Yöntemlerine Uygulanması, Veri Ön İşleme ve Öznitelik Mühendisliğinin, İstanbul, Duvar Yayınları, 33-54; 2023.
67. Park K, Chae M, Cho JH. Image pre-processing method of machine learning for edge detection with image signal processor enhancement. *Micromachines*. 2021;12(1):1-13
68. Dedetürk BA, Taşdemir K, Bakır Güngör B. Cilt Kanseri Görüntü Sınıflandırması için Görüntü Ön İşlemenin Evrimsel Sinir Ağları Performansı Üzerindeki Etkileri. *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi*. 2022;38(2):190-200.
69. Dursun ÖF, Türkmen Hİ. Süper Piksel Tabanlı Otomatik Kanlı Bölge Tespit Sistemi. *International J Advances in Engineering and Pure Sciences*. 2021;33(1):28-38.
70. Menteşe E, Hançer E. Histopatoloji görüntülerde derin öğrenme yöntemleri ile çekirdek segmentasyonu. *Avrupa Bilim ve Teknoloji Dergisi*. 2020;Özel Sayı:95-102.
71. Ball DL, Fisher RJ, Burmeister BH, Poulsen MG, Graham PH, Penniment MG, et al. The complex relationship between lung tumor volume and survival in patients with non-small cell lung cancer treated by definitive radiotherapy: a prospective, observational proGonzalez RC. *Digital image processing: Pearson Edu India*; 755-778:2009.
72. Sezgin M, Sankur BI. Survey over image thresholding techniques and quantitative performance evaluation. *J Electron Imaging*. 2004;13(1):146-68.
73. Bozkurt F. A deep and handcrafted features- based framework for diagnosis of COVID- 19 from chest x- ray images. *Concurr Comput*. 2022;34(5):e6725.

74. Bozkurt F. A study on CNN based transfer learning for recognition of flower species. *Avrupa Bilim ve Teknoloji Dergisi*. 2022(32):883-90.
75. Idris K, Bhoite S. Applications of machine learning for prediction of liver disease. *Int J Comput Appl Technol Res*. 2019;8(9):394-96.
76. Ge G, Shi Z, Zhu Y, Yang X, Hao Y. Land use/cover classification in an arid desert-oasis mosaic landscape of China using remote sensed imagery: Performance assessment of four machine learning algorithms. *Glob Ecol Conserver*. 2020;22:e00971.
77. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32.
78. Okumus H, Nuroglu FM. A random forest-based approach for fault location detection in distribution systems. *Electr Eng*. 2021;103(1):257-64.
79. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*; 2016;785 - 794
80. Ahmed NS, Huynh N, Gassman S, Mullen R, Pierce C, Chen Y. Predicting pavement structural condition using machine learning methods. *Sustainability*. 2022;14(14):1-16.
81. W Ahmed NS, Huynh N, Gassman S, Mullen R, Pierce C, Chen Y. Predicting pavement structural condition using machine learning methods. *Sustainability*. 2022;14(14):1-16
82. Dorogush AV, Ershov V, Gulin A. CatBoost: Gradient boosting with categorical features support. 2018, arXiv:1810.11363:1-7
83. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*. 10th ed. Clifornia, Curran Associates2018;1-11.
84. Yue S, Li P, Hao P. SVM classification: Its contents and challenges. *Appl. Math. J. Chinese Univ. Ser. B*. 2003;18:332-42.
85. İbm C. Spss I. Statistics for windows. Version 21. Armonk, Ny: IBM Corp. 2012.
86. Huang Y, Liu Z, He L, Chen X, Pan D, Ma Z, et al. Radiomics signature: a potential biomarker for the prediction of disease-free survival in early-stage (I or II) non—small cell lung cancer. *Radiology*. 2016;281(3):947-57.
87. Ball DL, Fisher RJ, Burmeister BH, Poulsen MG, Graham PH, Penniment MG, et al. The complex relationship between lung tumor volume and survival in patients with non-small cell lung cancer treated by definitive radiotherapy: a prospective,

- observational prognostic factor study of the Trans-Tasman Radiation Oncology Group (TROG 99.05). *Radiother Oncol.* 2013;106(3):305-11.
88. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Eur Radiol.* 2013;26:1045-057.
89. Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, et al. A radiogenomic dataset of non-small cell lung cancer. *Sci Data.* 2018;5(1):1-9.
90. Le NQK, Kha QH, Nguyen VH, Chen Y-C, Cheng S-J, Chen C-Y. Machine learning-based radiomics signatures for EGFR and KRAS mutations prediction in non-small-cell lung cancer. *Int J Mol Sci.* 2021;22(17):1-14.



EKLER

EK-1. Özgeçmiş



EK-2. Etik Kurul Onayı

