BERFU MERGEN
KANSER HÜCRE HATTI İLE İLAÇ DUYARLILIK İLİŞKİSİNİN AÇIK KAYNAKLI VERİLER KULLANILARAK ANALİZİ
2026



BURSA ULUDAĞ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

KANSER HÜCRE HATTI İLE İLAÇ DUYARLILIK İLİŞKİSİNİN AÇIK KAYNAKLI VERİLER KULLANILARAK ANALİZİ

BERFU MERGEN

TRANSLASYONEL TIP ANABİLİM DALI

DOKTORA TEZİ

Bursa 2026

T.C.
BURSA ULUDAĞ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
TRANSLASYONEL TIP ANA BİLİM DALI

KANSER HÜCRE HATTI İLE İLAÇ DUYARLILIK İLİŞKİSİNİN AÇIK KAYNAKLI VERİLER KULLANILARAK ANALİZİ

Berfu MERGEN

DOKTORA TEZİ

DANIŞMAN
Doç. Dr. Gıyasettin ÖZCAN

BURSA-2026

T.C.
BURSA ULUDAĞ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

ETİK BEYANI

Doktora tezi olarak sunduğum "**Kanser Hücre Hattı ile İlaç Duyarlılık İlişkisinin Açık Kaynaklı Veriler Kullanılarak Analizi**" adlı çalışmanın, proje safhasından sonuçlanmasına kadar geçen bütün süreçlerde bilimsel etik kurallarına uygun bir şekilde hazırlandığını ve yararlandığım eserlerin kaynaklar bölümünde gösterilenlerden oluştuğunu belirtir ve beyan ederim.

Berfu MERGEN
Tarih ve İmza
02/01/2026

TEZ KONTROL ve BEYAN FORMU

02/01/2026

Adı Soyadı: Berfu MERGEN

Anabilim Dalı: Translasyonel Tıp Ana Bilim Dalı

Tez Konusu: Kanser Hücre Hattı ile İlaç Duyarlılık İlişkisinin Açık Kaynaklı Veriler Kullanılarak Analizi

ÖZELLİKLER	UYGUNDUR	UYGUN DEĞİLDİR	AÇIKLAMA
Tezin Boyutları	■	❑	
Dış Kapak Sayfası	■	❑	
İç Kapak Sayfası	■	❑	
Kabul Onay Sayfası	■	❑	
Sayfa Düzeni	■	❑	
İçindekiler Sayfası	■	❑	
Yazı Karakteri	■	❑	
Satır Aralıkları	■	❑	
Başlıklar	■	❑	
Sayfa Numaraları	■	❑	
Eklerin Yerleştirilmesi	■	❑	
Tabloların Yerleştirilmesi	■	❑	
Kaynaklar	■	❑	

DANIŞMAN ONAYI

Doç.Dr. Gıyasettin ÖZCAN

İmza:

İÇİNDEKİLER

ETİK BEYAN ..... **I**
KABUL ONAY SAYFASI ..... Hata! Yer işareti tanımlanmamış.
TEZ KONTROL ve BEYAN FORMU ..... **III**
İÇİNDEKİLER ..... **IV**
TÜRKÇE ÖZET ..... **VI**
İNGİLİZCE ÖZET ..... **VII**
TEZ KONUSUNUN KÜRESEL SÜRDÜRÜLEBİLİR KALKINMA HEDEFLERİ İLE İLİŞKİSİ ..... **VIII**
1.GİRİŞ **1**
2.GENEL BİLGİLER **3**
2.1. Kanser Araştırmalarında Açık Veri Kaynakları 3
2.1.1. The Cancer Genome Atlas (TCGA) 4
2.1.2. International Cancer Genome Consortium (ICGC) 4
2.1.3. Cancer Cell Line Encyclopedia (CCLE) 5
2.1.4. Catalogue of Somatic Mutations in Cancer (COSMIC) 5
2.1.5. Genomic Data Commons (GDC) 5
2.1.6. Oncomine 6
2.1.7. The Human Protein Atlas (HPA) 6
2.1.8. BioGRID (Biological General Repository for Interaction Datasets) 7
2.1.9. DGIdb (Drug–Gene Interaction Database) 7
2.1.10. GDSC (Genomics of Drug Sensitivity in Cancer) 7
2.2. Kanser Hücre Hatları 8
2.3. Kanserde İlaç Duyarlılığının Mekanizmaları 11
2.4. Sağlık Alanında Makine Öğrenmesi 14
2.5. İlaç Duyarlılığının Tahmin Edilmesinde Makine Öğrenmesi 16
2.6. Farmakogenomiğin Ortaya Çıkışı 17
2.7. Mutasyon Verisi ve İlaç Yanıtı Tahminindeki Önemi 18
2.8. Kopya Sayısı Değişimi Verisi ve İlaç Yanıtı Tahminindeki Önemi 20
2.9. Moleküler Biyoloji, Genomik ve Hesaplama Teknolojilerindeki İlerlemeler 22
2.10. Kişiselleştirilmiş Tıp 23
2.11. Onkolojide Sürücü Mutasyonlar ve Yaklaşımlar 23
2.12. İlaç-İlaç Etkileşimleri ve Hesaplamalı Yaklaşımlar 24
2.13. Veri Entegrasyonu ve Makine Öğrenmesi Modelleri 25
3.GEREÇ VE YÖNTEM **28**
3.1. Çalışmanın Birinci Aşaması ile İlgili Gereç ve Yöntem 28
3.1.1. Veri toplama 28
3.1.2. Lung Small Cell Carcinoma verisi 29
3.1.3. Makine öğrenmesi için veri seti oluşturma 30
3.1.4. Makine öğrenmesi uygulaması 31
3.1.5. Makine öğrenmesi modelleme 31
3.1.6. İdeal model parametre belirlenmesi 32
3.2. Çalışmanın İkinci Aşaması ile İlgili Gereç ve Yöntem 34
3.2.1. Veri setlerinin birleştirilmesi ve ön işleme süreci 34
3.2.2. Açık erişimli veri kaynaklarının toplanması ve entegrasyonu ve

GDSC verilerinin toplanması 34
3.2.3. GDSC ve COSMIC verilerinin entegrasyonu 35
3.2.4. DGIdb'nin entegrasyonu 36
3.2.5. COSMIC genom tarama mutasyonlarının entegrasyonu 37
3.2.6. Veri ön işleme ve özellik seçimi 37
3.2.7. Kategorik verilerin kodlanması ve eksik alanların düzenlenmesi 38
3.2.8. Eksik verilerin tamamlanması ve yanlılığın azaltılması 38
3.2.9. Drug–Gene etkileşimlerinin keşifsel incelemesi 39
3.2.10. İlaç ve genlerin kümeleme analizleri.....39
3.2.11. Modelleme öncesi özellik seçiminde optimizasyon.....40
3.2.12. Son veri seti ve performans ölçütleri 40
3.2.13. Makine öğrenmesi modellerinin oluşturulması.....41
3.2.14. Veri setinin bölünmesi 41
3.2.15. Özellik seçimi ve algoritmaların uygulanması.....41
4.BULGULAR 43
4.1. Çalışmanın Birinci Aşaması ile İlgili Bulgular 43
4.1.1. Mutasyon yük veri üretmenin zorluğu.....43
4.1.2. Makine öğrenme algoritma ile tahminler 44
4.1.3. Mutasyon türlerinin etkisi ve model karmaşıklık analizi.....46
4.2. Çalışmanın 2. Aşaması ile İlgili Bulgular 50
4.2.1. Veri entegrasyonu ve deneysel tasarım.....50
4.2.2. İlaç–Gen etkileşimlerinin incelenmesi 50
4.2.2.1. Etkileşim sıklığı analizi 51
4.2.2.2. İlaç duyarlılık analizi 52
4.2.3. Kümeleme ve korelasyon analizleri 54
4.2.3.1 Hiyerarşik kümeleme bulguları 54
4.2.3.2. Korelasyon analizine ilişkin bulgular 56
4.2.4. Makine öğrenmesi modellerinin performans analizi.....59
4.2.5. Ablasyon analizi.....61
4.2.6. Bağımsız CTRP–CCLE Farmakogenomik Verisi Üzerinde Değerlendirme.....64
5.TARTIŞMA VE SONUÇ.....66
6.KAYNAKLAR 85
7.SİMGELER VE KISALTMALAR 95
8. TEŞEKKÜR 97
9.ÖZGEÇMİŞ.....98

TÜRKÇE ÖZET

Kanser dokuları, biriken mutasyonlar nedeniyle genetik olarak dinamik bir yapı sergilemekte olup, yeni mutasyonların ortaya çıkması kanser ilaçlarının etkinliğini değiştirebilmektedir. Bu çalışmanın amacı, genetik mutasyon profilleri ile ilaç duyarlılığı arasındaki ilişkiyi analiz etmek ve bu ilişkiyi kullanarak kanser hücrelerinin ilaç yanıtını tahmin edebilen bir makine öğrenmesi modeli geliştirmektir.

Çalışmanın ilk aşamasında, küçük hücreli akciğer kanserinde ilaç–mutasyon etkileşimleri incelenmiştir. Hücre hatları, mutasyon yükü, doku bilgisi ve IC50 değerleri tek bir veri kümesinde birleştirilmiştir. Oluşturulan veri kümesi, mutasyon yüküne bağlı olarak ilaç duyarlılığını tahmin etmek amacıyla makine öğrenmesi analizlerinde kullanılmıştır. Çalışmanın ikinci aşamasında ise daha geniş kapsamda, tüm kanser hücre hatlarına ait ilaç duyarlılığı verileri, somatik mutasyon verileri ve gen–ilaç etkileşim bilgilerini içeren farklı veri tabanları entegre edilmiştir.

Deneysel sürecin ilk aşamasında, mutasyon türü, sayısı, biyolojik etkisi ve varyant alel frekansı gibi özellikler sayısallaştırılarak her hücre hattı için tanımlayıcı öznitelikler elde edilmiştir. Yapılan analizlerde, ilaç duyarlılığını en yüksek doğrulukla tahmin eden algoritmanın XGBoost olduğu belirlenmiştir. XGBoost, R^2, RMSE ve MAE performans ölçütleri açısından diğer algoritmalara kıyasla daha başarılı sonuçlar vermiştir. Mutasyon yükü özellikleri kullanılarak küçük hücreli akciğer kanserinde IC50 değeri anlamlı düzeyde tahmin edilmiştir ($R^2 \approx 0{,}60$). Çalışmanın ikinci aşamasında ise LightGBM Regressor en yüksek performansı göstermiştir. Model, $R^2 = 0{,}9120$ değeri ile öne çıkarken aynı zamanda en düşük MSE (0,1371) ve MAE (0,1576) değerlerini üretmiştir.

Geliştirilen modeller, deneysel çalışmalara yön verebilecek öngörüsel bir karar destek yaklaşımı sunmaktadır. Elde edilen bulgular, genetik mutasyonların ilaç yanıtındaki belirleyici rolünü ortaya koymakta ve kişiselleştirilmiş kanser tedavisine katkı sağlayabilecek güçlü bir çerçeve sunmaktadır.

Anahtar Kelimeler: İlaç duyarlılığı, Z-skoru, çoklu veri kaynaklarının entegrasyonu, veri işleme, makine öğrenmesi modelleri

İNGİLİZCE ÖZET

ANALYSIS OF THE RELATIONSHIP BETWEEN CANCER CELL LINES AND DRUG SENSITIVITY USING OPEN SOURCE DATA

Cancer tissues exhibit a genetically dynamic structure due to the accumulation of mutations, and the emergence of new mutations may alter the efficacy of anticancer drugs. The aim of this study is to analyze the relationship between genetic mutation profiles and drug sensitivity and to develop a machine learning model capable of predicting drug response in cancer cells based on this relationship.

In the first phase of the study, drug–mutation interactions were investigated in small cell lung cancer. Cell line data, mutation burden, tissue information, and IC50 values were integrated into a single dataset. The resulting dataset was used in machine learning analyses to predict drug sensitivity based on mutation burden. In the second phase, a broader approach was adopted by integrating drug sensitivity data from all cancer cell lines, somatic mutation profiles, and gene–drug interaction information obtained from multiple databases.

During the first stage of the experimental process, biological features such as mutation type, number, functional impact, and variant allele frequency were quantified to generate descriptive features for each cell line. The analyses revealed that XGBoost achieved the highest predictive accuracy for drug sensitivity. XGBoost outperformed other algorithms in terms of R^2, RMSE, and MAE performance metrics. Using mutation burden features, the IC50 value in small cell lung cancer was predicted at a statistically significant level ($R^2 \approx 0,60$). In the second phase, the LightGBM Regressor demonstrated the best performance, achieving the highest R^2 value (0,9120) while producing the lowest MSE (0,1371) and MAE (0,1576).

The developed models provide a predictive decision-support framework that may guide experimental studies. The findings highlight the critical role of genetic mutations in determining drug response and present a robust framework that may contribute to personalized cancer therapy.

Keywords: Drug sensitivity, Z-score, integration of multiple data sources, data preprocessing, machine learning models

BUÜ SAĞLIK BİLİMLERİ ENSTİTÜSÜ
TEZ KONUSUNUN KÜRESEL SÜRDÜRÜLEBİLİR KALKINMA HEDEFLERİ İLE İLİŞKİSİ



Doktora tezi olarak sunduğum **"**Kanser Hücre Hattı ile İlaç Duyarlılık İlişkisinin Açık Kaynaklı Veriler Kullanılarak Analizi**"** başlıklı tez **3.** Küresel Sürdürülebilir Kalkınma Hedefleri ile ilişkilidir.

1. GİRİŞ

Hassas onkoloji, bir hastanın tümörünün özelliklerine göre kişiye özel tedaviler sunmayı amaçlar. Bu hedef daha fazla veri ve daha iyi hesaplama modelleri ile ilaç yanıtını artan doğrulukla tahmin edebileceğimiz fikrine dayanmaktadır. Nitekim son zamanlardaki iki eğilim bunu desteklemektedir. İlk olarak yüksek verimli teknolojiler (high-throughput technologies) genetik veri miktarını önemli ölçüde artırmıştır. Gen ekspresyonu ve mutasyon profilleri gibi omik verilerin artmış olması artık potansiyel ilaç yanıtı tahmin edicileri için incelenebilir düzeye gelmiştir. Tahmin hedef tarafında, klinik sonuçlar hala sınırlı olup, hasta ilaç yanıtını değişen doğrulukla taklit eden tarama sonuçları klinik öncesi modellerde bol miktarda bulunmaktadır. İkinci olarak derin öğrenme, veriden yararlanmak için doğal bir teknik olarak ortaya çıkmıştır. Geleneksel istatistiksel modellerle karşılaştırıldığında, yüksek sinir ağlarının kapasitesi, moleküler ve ilaç özellikleri arasındaki karmaşık etkileşimleri daha iyi yakalamalarını sağlamaktadır.

Bu faktörlerin bir araya gelmesi ilaç tepkisini tahmin etmek için hesaplamalı modellerin yeni bir neslinin ortaya çıkmasına neden oldu. The National Cancer Institute 60 (NCI60) insan tümör hücre hattı veritabanı antikanser ilaç taraması için en eski kaynaklardan biridir. Zengin moleküler karakterizasyon verileri araştırmacıların farklı analizlerin yöntemlerini tahmin gücünü karşılaştırmasına olanak sağlamıştır. Örneğin, gen ifadesi, protein ve mikroRNA verilerinin çokluğu hem tekli hem de eşleştirilmiş ilaç yanıtını tahmin etmek için etkili özellik türleridir.

İlaç yanıtı tahmini, sadece ilaç ve hücre hattı analizinin ötesine geçmiş, bütünleştirici modeller dahil edilerek hem ilacı hem de hücre hattını girdi özellikleri olarak alan bir yaklaşıma dönüşmüştür. Gittikçe artan bilimsel ilgi hesaplama yaklaşımlarına daha fazla ilham vermiştir. Kanser hücre hatları ile ilgili kozmik veri miktarı, son yıllarda hızla artmaktadır. İnsan genetik çalışmaları sonucu elde edilen verinin karmaşıklığı ve boyutunun büyüklüğü, verinin doğru bir şekilde analizini güçleştirmektedir. Böylece hem bireyin tedavisi hem de toplum refahı açısından başta sağlık ve ekonomik sonuçlar olmak üzere olumsuz etkilere sebep olmaktadır.

Çalışmamız sürecinde açık kaynaklı olan ve literatürde yer alan ilaç-gen-mutasyon-hastalık veri tabanları araştırılmıştır. Elde edilen bu veri tabanları

akademik camiada kabul gören kurumlar tarafından sunulan hali hazırda açık kaynaklıdır. Bunun yanı sıra, bu veriler dergi etki faktörü üst seviye olan dergilerde araştırma makalesi olarak yayınlanmıştır ve uluslararası araştırmalarda kullanılmak üzere herkesin erişimine açıktır. Bu verilere erişim, analiz ve bilimsel araştırma süreçleri herhangi bir etik süreç gerektirmemektedir.

Elde edilen verilerin bellekte yüksek alana sahip olduğu görülmektedir ve elle işlenmesi mümkün değildir. Öte yandan hesaplama süreçlerinin bilgisayar yardımı ile yapılması dahi ciddi zorluklar içermektedir. Zira verilerden doğru örüntülerin elde edilmesi için bilgisayar tarafından hesaplanamayacak kadar çok sayıda ihtimal vardır. Bu nedenle verilerde filtreleme, dizinleme, örüntü bulma süreçleri yürütülmüştür. Çalışmada gen-mutasyon-hastalık-ilaç etkileşimi konusunda tekrar eden örüntüler yakalanması hedeflenmiştir. Bu doğrultuda veri kümeleri arasında anlamlı ilişkiler bulan yapay öğrenme algoritmaları kodlama kullanarak deneysel olarak geliştirilebilecektir.

Çalışmanın amacı gen mutasyonları, gen ilaç etkileşimleri, kanser hücre hatları ve ilaç duyarlılığı verilerinin analiz edilerek yeni örüntüler bulmaktır. Araştırmamızda temel aldığımız varsayım; hücre hattı ilaç yanıtının birden fazla değişkene dayanması, hücrenin moleküler özelliklerinin bu yanıtların oluşmasında önemli rol oynamasına bağlı olarak bu nitelikleri temsil eden genetik veri tiplerinin kullanılmasıyla ilaç yanıtının modellenebilmesidir. Bu örüntüler ve yeni veri analizi yöntemleri bulunması sonucu daha hızlı, daha isabetli ve daha ekonomik tedavilere ulaşmak mümkün olabilecektir.

2. GENEL BİLGİLER

2.1. Kanser Araştırmalarında Açık Veri Kaynakları

Kanser araştırmalarında, araştırmacıların verileri maliyet etkin bir şekilde göndermesine, paylaşmasına, erişmesine ve analiz etmesine yardımcı olarak kanser araştırmalarını hızlandıran güvenli açık veri kaynaklarının kullanımı her geçen gün artmaktadır. Ayrıca, kanseri önleme, teşhis etme ve tedavi etme yolunda ilerlemeyi hızlandıracak keşifler yapmaya yardımcı olmak amacıyla açık veri kaynakları gelecek vadetmektedir.

2022 yılında yaklaşık on milyon ölüm, yaklaşık yirmi milyon yeni vaka ve 2050 yılına kadar otuzbeş milyon yeni vakanın tahmin edildiği kanser, hâlâ dünya çapında ikinci en yüksek ölüm nedenidir ve oldukça ciddi bir halk sağlığı sorunudur. Tahminlere göre dokuz erkekten biri ve on iki kadından biri kanserden ölecek ve beş erkek ve kadından biri hayatının bir noktasında bu hastalığa yakalanacaktır (Bray ve ark., 2024; Ferlay ve ark., 2024). Genel olarak, önde gelen ölüm nedenleri kolorektal kanser ve akciğer kanseridir. Bu sıralamayı mide, karaciğer ve meme kanserleri takip etmektedir. Kanser görülme sıklığının artmasının ana nedenleri akciğer, meme, kolorektal, prostat ve mide kanserleridir (World Health Organization [WHO], 2024).

Kanser hastalarının tedavisi histolojiye ve tümör evresine bağlıdır. Hastalığın erken evrelerinde teşhis edilen hastalar için cerrahi tedavi tercih edilen tedavi yöntemi olmakla birlikte kemoterapi ve radyoterapi ileri aşamalar için standart tedavilerdir. Ancak daha az toksik ve daha etkili kemoterapi tedavilerinin geliştirilmesine rağmen kesin tedavi yöntemlerinin olmayışı, hedefler ve tedavilere karşı devam eden tümör direnci, hassas onkolojinin ortaya çıkış ihiyacını acil hale getirmiştir. Bu yaklaşım onkojenik moleküler değişiklikleri ortadan kaldırmayı amaçlayan tedaviler, hastaların seçimi ve bu değişikliklere sahip hastaların malignitelerine ilaç uygulayan yeni temellere dayanmaktadır (Lawson, Kessenbrock, Davis, Pervolarakis, & Werb, 2018). Son zamanlarda immünoterapi melanom ve akciğer kanseri de dahil olmak üzere çeşitli kanser türleri için cesaret verici sonuçlarından dolayı en cazip tedavi seçeneği haline gelmiştir (Baik ve ark., 2017). Bununla birlikte, kaydedilen kayda değer ilerlemelere rağmen tedavi yöntemlerinde

yapılan değişiklikler, birçok kanser hastasının nüks ve metastaza maruz kalmasına engel olamamıştır (Ganesh, & Massagué , 2021).

Kanser araştırmalarında açık veri kaynakları, kanserin biyolojisini, genetik temellerini, tedavi yöntemlerini ve çeşitli klinik özelliklerini daha derinlemesine öğrenmeyi amaçlayan çok çeşitli veriler sağlar. Bu kaynaklar, genetik ve moleküler düzeyde kanserin mekanizmalarını anlamak, tedaviye yönelik daha hedeflenmiş yaklaşımlar geliştirmek ve biyobelirteçleri tanımlamak için kritik öneme sahiptir. Bu kaynakların en yaygın kullanılanları şu şekilde sıralanabilir (The Cancer Genome Atlas Research Network, 2013).

2.1.1. The Cancer Genome Atlas (TCGA)

TCGA, Amerikan Kanser Enstitüsü (NCI) ve Ulusal Genomik Araştırmalar Enstitüsü (NHGRI) tarafından desteklenen bir projedir. 33 farklı kanser türüne ait genetik ve klinik veriler içerir. Bu kaynak, kanserin genetik yapısını keşfetmek için geniş çaplı biyolojik örnekler ve veriler sunar. Veri Türleri; genomik veriler (DNA dizilimleri, somatik mutasyonlar, gen ekspresyonu, DNA metilasyonu), çeşitli kanser türlerine ait klinik veriler (yaş, cinsiyet, tedavi türleri vb.), histopatolojik veriler, proteomik ve mikrobiyom verileridir. TCGA, kanserin genetik ve biyolojik çeşitliliği üzerinde derinlemesine analiz yapılmasını sağlar. Bu veriler, kanserin biyolojik temellerini anlamada ve daha hedeflenmiş tedavi stratejilerinin geliştirilmesinde kullanılır (https://portal.gdc.cancer.gov/). (The Cancer Genome Atlas Research Network (2013).

2.1.2. International Cancer Genome Consortium (ICGC)

ICGC, dünya çapında kanserin genomik temellerini araştırmayı amaçlayan bir uluslararası işbirliği projesidir. 50'den fazla araştırma grubunun katılımıyla, kanser türlerine göre detaylı genetik veriler toplar ve bu verileri kamuya açık hale getirir. Veri türleri; genomik veriler (somatik mutasyonlar, kopya numarası değişiklikleri, mikrosatellit instabilitesi vb.), epigenetik veriler (DNA metilasyonu), transkriptomik ve proteomik verilerdir. Genetik çeşitlilik üzerine kapsamlı çalışmalar yapılmasına olanak tanır. Kanserin evrimini, biyolojik işlevlerini ve potansiyel tedavi hedeflerini

anlamaya yönelik önemli veriler sağlar (https://icgc.org/). (International Cancer Genome Consortium (2010).

2.1.3. Cancer Cell Line Encyclopedia (CCLE)

CCLE, kanser hücre hatlarına ait genetik ve biyolojik verileri içeren geniş bir veri tabanıdır. Bu platform, kanser tedavisi için hücre hatlarını kullanarak çeşitli tedavi stratejilerinin etkinliğini inceleyen araştırmalar için oldukça önemlidir. Veri türleri; kanser hücre hatlarının genetik profilleri, gen ekspresyonu ve protein düzeyleri, farmakolojik duyarlılık verileri (kanser tedavisi ile ilişkili), DNA dizilimi, CNV (copy number variations) ve mutasyon verileridir. Hücre hatlarının moleküler profilleri, kanserin biyolojisini anlamada ve tedaviye yönelik hedeflerin belirlenmesinde kullanılır. Kanser tedavilerine duyarlılığı test etmek için geniş veri setleri sağlar (https://portals.broadinstitute.org/ccle) (Barretina ve ark., 2012).

2.1.4. Catalogue of Somatic Mutations in Cancer (COSMIC)

COSMIC, kanserle ilişkili somatik mutasyonların bulunduğu büyük bir veritabanıdır. Bu kaynak, kanserin genetik temelleri hakkında çok önemli bilgiler sunar, özellikle somatik mutasyonların hangi kanser türlerinde yaygın olduğuna dair derinlemesine bilgi verir. Veri türleri; somatik mutasyon verileri (genetik varyasyonlar, mutasyonlar, Fokal Kopya Sayısı değişiklikleri), genetik mutasyonların kanser türlerine göre dağılımı, kanser türlerine özgü genetik profil bilgileridir. Somatik mutasyonların kanserin gelişimine etkilerini anlamak için önemli bir kaynaktır. Araştırmacılara, kanserin moleküler temelini daha iyi anlamak için veri sağlar (https://cancer.sanger.ac.uk/cosmic) (Forbes ve ark., 2017).

2.1.5. Genomic Data Commons (GDC)

GDC, kanser genomik verilerinin toplanması, depolanması ve analiz edilmesi için bir platformdur. TCGA ve diğer projeler tarafından sağlanan verileri içerir. Bu veri kümesi, kanser araştırmalarına yönelik çok çeşitli genetik, epigenetik ve klinik

bilgileri sunar. Veri türleri; genomik veriler (DNA dizilim verileri, mutasyonlar, CNV, gen ekspresyonu, DNA metilasyonu), klinik veriler (yaş, cinsiyet, tedavi türü, hastalık evresi), histopatolojik veriler ve görüntüleme verileridir. GDC, TCGA ve diğer kaynaklardan gelen verilerin bir araya getirildiği merkezi bir veri deposudur. Kanser araştırmalarına odaklanan derinlemesine analizlere olanak tanır (https://gdc.cancer.gov/) (Jensen ve ark., 2017).

2.1.6. Oncomine

Oncomine, kanserle ilişkili gen ekspresyon verilerini sunan bir veri tabanıdır. Kanser hücre hatlarının, tümörlerin ve kanserle ilişkili moleküllerin ekspresyon seviyelerini inceleyerek kanserin biyolojik süreçlerini anlamaya yardımcı olur. Veri türleri; gen ekspresyon verileri, somatik mutasyon verileri, onkogen ve tümör baskılayıcı genlerin ekspresyon seviyeleri, klinik verilerle ilişkilendirilmiş genetik verilerdir. Kanserin moleküler mekanizmalarını anlamak için geniş bir gen ekspresyon veri kümesi sağlar. Oncomine, araştırmacılara kanser türleri ve genetik mutasyonlar arasındaki ilişkiyi keşfetme fırsatı sunar (https://www.oncomine.org/) (Rhodes ve ark., 2007).

2.1.7. The Human Protein Atlas (HPA)

HPA, insan vücudundaki proteinlerin ekspresyonunu ve dağılımını araştıran bir kaynaktır. Bu platform, kanserin gelişimiyle ilişkili proteinlerin özelliklerini incelemek için kullanılır. Veri türleri; protein ekspresyon verileri (farklı organlar ve dokularda), kanserle ilişkili proteinlerin ekspresyon profilleri, kanser hücrelerinde proteinlerin dağılımıdır. Kanserin biyolojik temellerini anlamak ve tedavi hedefleri geliştirmek için faydalıdır. İnsan proteomunun detayıyla, kanser tedavisinde hedeflerin belirlenmesinde yardımcı olur (https://www.proteinatlas.org/) (Thul ve ark., 2017).

2.1.8. BioGRID (Biological General Repository for Interaction Datasets)

BioGRID, biyolojik etkileşim verilerini içeren bir veritabanıdır. Kanser araştırmalarında, genetik ve protein etkileşimlerinin belirlenmesi kanser biyolojisini anlamada kritik öneme sahiptir. Veri türleri; protein-protein etkileşimleri, genetik etkileşimler, etkileşimlerin zaman ve ortam bağımlılıklarıdır. Kanserin moleküler etkileşim ağlarını çözümlemek ve biyolojik süreçleri anlamak için kullanılır. Kanserle ilişkili biyolojik etkileşim ağlarını haritalamak için bir kaynaktır (https://thebiogrid.org/). Bu açık veri kaynakları kanser araştırmalarında büyük bir bilgi birikimi sağlar. Araştırmacılar bu platformlar aracılığıyla kanserin genetik, biyolojik ve klinik yönlerini inceleyerek daha etkili tedavi yöntemleri geliştirebilir, tedaviye dair biyobelirteçler keşfedebilir ve kanserin tedaviye olan yanıtını tahmin edebilirler (The Cancer Genome Atlas Research Network, 2013) (Oughtred ve ark., 2021).

2.1.9. DGIdb (Drug–Gene Interaction Database)

Genler ile ilaçlar arasındaki etkileşimleri sistematik olarak derleyen açık erişimli bir biyoinformatik veri tabanıdır. Bu veri tabanı; ilaç hedefleri, inhibitör ve aktivatör ilişkileri, klinik olarak anlamlı gen–ilaç etkileşimleri ve potansiyel terapötik hedefler hakkında küratörlü ve entegre edilmiş bilgiler sunmaktadır. DGIdb, farklı biyomedikal kaynaklardan elde edilen gen–ilaç etkileşim verilerini bir araya getirerek analizlerde, hedefe yönelik tedavi yaklaşımlarının belirlenmesinde ve ilaç yeniden konumlandırma çalışmalarında yaygın olarak kullanılmaktadır (https://www.dgidb.org) (Freshour ve ark., 2021).

2.1.10. GDSC (Genomics of Drug Sensitivity in Cancer)

Kanser hücre hatlarında ilaç yanıtı ile genomik özellikler arasındaki ilişkiyi inceleyen kapsamlı bir farmakogenomik veri tabanıdır. Veri tabanı; yüzlerce antikanser ilaca ait duyarlılık ölçümleri (IC50), gen ekspresyonu, mutasyon, kopya sayısı değişimleri ve diğer moleküler profilleri içermektedir. GDSC, biyobelirteç

keşfi, ilaç duyarlılığı modelleme ve hedefe yönelik tedavi stratejilerinin geliştirilmesi amacıyla yaygın olarak kullanılmaktadır (https://www.cancerrxgene.org) (Yang ve ark., 2013).

2.2. Kanser Hücre Hatları

Kanser türlerine yönelik etkin farmakolojik çözümlerin keşfinin çeşitli zorluklar ve yüksek ekonomik maliyetler barındırması, ilaç sanayi ve araştırmacıları yeni yaklaşımlarla çözüm getirme yoluna sevketmiştir. Bilgisayar teknolojilerindeki ilerlemelerin biyolojik veriyi daha erişilebilir ve ucuz kılmasıyla beraber kanser gibi karmaşık hastalıkların tedavisi için hassas tıp uygulamalarının önü açılmıştır. Tümör dokusunun modellenebilmesi için, yaygın olarak hızlı sonuç alınabilmesi ve çok daha az maliyetli oluşu nedeniyle hücre hatları kullanılmaktadır. Genetik paneller, ilaçların hücre hatları üzerindeki etkisini araştırmayı ve kanser hücrelerine has moleküler özelliklerin ortaya konulmasını hedeflemektedir. Hesaplamalı algoritmik yaklaşımlar ile moleküler özellik verisi içindeki örüntüler belirlenebilmekte ve deneysel verisi olmayan benzer özelliğe sahip başka hücre hatları de için ilaç yanıtı tahmini yapılabilmektedir. Yakın gelecekte bu yaklaşımla hastalar için daha doğru tedavi seçenekleri daha az iş gücü ve maliyet ile belirlenebilecektir.

Kanser hücre hatları, ilaç keşfi ve kanser araştırmalarında sıklıkla kullanılan in vitro model sistemlerdir (Masters, 2000). Araştırma için sonsuz miktarda biyolojik materyal sağlama konusundaki eşsiz yetenekleri, kullanımlarının ana nedenidir (Neve ve ark., 2006).

Yeni bir hücre dizisi oluşturma süreci son derece karmaşıktır ve hala tam olarak anlaşılamamıştır. Herbir örnek için başarılı oluşum oranı düşük ve tahmini oldukça güçtür (Drexler, Matsuo, & MacLeod, 2000). Bir hücre dizisinin stabilitesinin, hücre döngüsünün düzenlenmesinde ve apoptoz yoluyla hücre ölümünde yer alan tüm biyolojik sistemlerden kaçarken in vivo hızla gelişebilen tümörlerin bir örneğiyle başladığı göz önüne alındığında, bu durum karmaşık görünebilmektedir (Asfari ve ark., 1992; Sramkoski ve ark., 1999;. Komatsu ve ark., 1991; Sudo, Kodama, Amagai, Yamamoto, & Kasai, 1983). Yeni bir hücre hattı oluşturulmasındaki zorluk ve rastlantısallığın birçok nedeni, in vivo (vücut içi) ile in vitro (laboratuvar) mikro

ortamları arasındaki belirgin farklılıkların (örneğin, büyüme faktörlerine olan bağımlılık, oksijen konsantrasyonu, stromal ve immün hücrelerle etkileşim gibi) dikkate alınmasıyla açıklanabilir. Bu durum, kronik miyeloid löseminin kronik evresinden sürekli bir hücre hattı oluşturmanın neden mümkün olmadığını ortaya koymaktadır; hastalık in vivo'da lösemi hücrelerinin yüksek oranda çoğalmasıyla karakterize edilirken, aynı hücreler in vitro'da sadece birkaç hafta hayatta kalabilmektedir (Drexler, Matsuo, & MacLeod, 2000). Kanser hücre hatlarının tarihsel gelişimi ve doğru kullanımının kronolojik olarak Tablo 1'de görüldüğü gibi listelenmesi bu konudaki çalışmaların anlaşılması açısından önemlidir (Mirabelli, Coppola, & Salvatore, 2019).

Tablo 1. Kanser hücre hatlarının tarihsel gelişimi ve doğru kullanımının kronolojik olarak tanımlanması

Tarih	Açıklama
1907	İlk in vitro hücre kültürü deneyleri: embriyonik dokuların laboratuvar ortamında büyütülmesi
1911	Rous sarkomundan türetilen tümör dokularının in vitro ortamda kültürü
1951	HeLa hücresinin (ilk insan kanser hücre hattı) stabilizasyonu ve karakterizasyonu
1956	Mikoplazmanın hücre kültürü için bir kontaminant (bulaş) etmeni olarak tanımlanması
1963	RAJI hücre hattının (ilk insan hematopoietik hücre hattı) oluşturulması
1970	K562 hücre hattının (kronik miyeloid lösemiden türetilen ilk hematopoietik hücre hattı) kurulması
1981	Hücre hatlarının çapraz kontaminasyonu ve yanlış tanımlanmasının araştırmalarda önemli bir sorun olduğunun belirlenmesi
1982	Ulusal Kanser Enstitüsü'nün (NCI) NCI60 antikanser tarama panelini başlatması
1986	NB4 hücre hattının stabilizasyonu ve karakterizasyonu: akut promyelositik lösemi çalışmaları için in vitro model sistem
1989	İnsan malign hücre hatlarının karakterizasyonu ve yayınlanmasına ilişkin kılavuzların yayımlanması
1990	Hücre hatlarının kanser araştırmalarında kullanımına yönelik UKCCCR (Birleşik Krallık Kanser Araştırma Kurumları Koordinasyon Komitesi) kılavuzlarının yayımlanması
2000 – 2012 – 2019	Hücre hatlarının genomik, transkriptomik ve proteomik karakterizasyonuna ilişkin kapsamlı veri setlerinin yayımlanması

Kanser biyolojisi, ilaç geliştirme ve özelleştirilmiş tedavi araştırmaları için tekrarlanabilir, standartlaştırılmış modeller sunan bu keşifler, kanser araştırmaları alanına önemli katkılarda bulunmuştur (Mirabelli, Coppola, & Salvatore, 2019). Hücre hatları, birincil insan dokusunun aksine, zaman içinde tekrar tekrar kullanılabilen değerli bir kaynağı temsil eder. Mutasyon içeriği ve mRNA

ekspresyonu açısından kolaylıkla karakterize edilebilirler (Liu, & Bodmer, 2006). Ayrıca, birincil insan dokusundan elde edilen verilerin yorumlanmasını etkileyebilecek stromal doku ile kontamine değildirler (Volchenboum ve ark., 2009). Hücre hatlarının orijinal olarak izole edildikleri tümörleri temsil ettiğine dair güçlü kanıtlar bulunmuştur (Douglas ve ark., 2004) ve orijinal dokuda bulunanlara benzer yapılar geliştirebilir (Willson, Bittner, Oberley, Meisner, & Weese, 1987).

Kanser hücre hatları, in vitro model sistemler olarak tümör biyolojisinin, genetik varyasyonların ve farmakolojik yanıtların incelenmesinde kritik öneme sahiptir. Özgün kanser türlerinden türetilmiş olan bu hücre hatları, laboratuvar ortamında sürekli çoğalabildikleri için moleküler onkoloji ve klinik öncesi araştırmalarda yaygın olarak kullanılmaktadır (Shoemaker, 2006; Weinstein ve ark., 1997).

Cancer Cell Line Encyclopedia (CCLE) projesi, 947'den fazla insan kanser hücre hattının genomik, transkriptomik, epigenetik ve farmakolojik profillerini sunarak, kanserin moleküler temellerinin daha iyi anlaşılmasına katkıda bulunmaktadır. Proje, kanser hücre hatlarının genetik mutasyonlarını, kopya sayısı varyasyonlarını ve ilaç duyarlılıklarını detaylı şekilde analiz etmiştir. Bu yaklaşım, tedavi yanıtlarının tahmin edilmesi ve bireyselleştirilmiş onkoloji stratejilerinin geliştirilmesi açısından önemli bir referans oluşturmuştur (Barretina ve ark., 2012; Ghandi ve ark., 2019). Genomics of Drug Sensitivity in Cancer (GDSC) projesi, farklı kanser hücre hatlarının genetik verilerini analiz ederek, çeşitli anti-kanser ajanlarına karşı duyarlılıklarını kapsamlı bir şekilde değerlendirmektedir. Bu veri tabanı, hücre hatlarının genetik profillerini ilaç yanıtlarıyla ilişkilendirerek, tedaviye yönelik biyobelirteçlerin keşfine olanak tanımaktadır. GDSC çalışmaları, kanser tedavilerinde bireyselleştirilmiş stratejilerin geliştirilmesi için kritik öneme sahiptir (Iorio ve ark., 2016; Yang ve ark., 2013).

The NCI-60 Human Tumor Cell Line PaneL, ABD Ulusal Kanser Enstitüsü (NCI) tarafından geliştirilen NCI-60 paneli, 60 farklı insan kanser hücre hattını içeren ve kanser ilaç taramalarında yaygın olarak kullanılan en erken model sistemlerden biridir. Bu panel, hücre hatlarının genetik profillemesi ve farmakolojik tepkilerinin incelenmesinde temel bir referans olarak öne çıkmaktadır. NCI-60, preklinik araştırmalarda ilaç etkinliğinin ve yanıtlarının belirlenmesinde kullanılmıştır (Shoemaker, 2006; The Cancer Genome Atlas Research Network,

2013).

Kanser hücre hatları, moleküler onkoloji ve genetik çalışmalarında temel model sistemler olarak, kanserin moleküler temellerinin aydınlatılmasında ve yeni tedavi stratejilerinin geliştirilmesinde vazgeçilmez araçlardır. CCLE, GDSC ve NCI-60 gibi veri tabanları, kanser hücre hatlarının kapsamlı genetik ve farmakolojik profillemesi sayesinde, kanser biyolojisinin derinlemesine analizini mümkün kılmaktadır. Bu kapsayıcı yaklaşımlar, hedefe yönelik tedavilerin belirlenmesi ve kişiselleştirilmiş onkoloji uygulamalarının geliştirilmesinde önemli rol oynamaktadır.

2.3. Kanserde İlaç Duyarlılığının Mekanizmaları

Kanser tedavisinde karşılaşılan en büyük problemlerden biri ilaç duyarlılığıdır. Kemoterapi, hedefe yönelik tedaviler ve immünoterapilere karşı gelişen direnç, tedavi etkinliğini azaltarak hastalık gelişimini olumsuz yönde etkileyebilmektedir (Holohan, Van Schaeybroeck, Longley, & Johnston, 2013). Tümör hücrelerinin doğrudan genetik ve epigenetik değişikliklerle veya tümör mikro çevresinin dinamikleri aracılığıyla geliştirdiği karmaşık biyolojik süreçler ilaç duyarlılığının en önemli sebepleridir (Vasan, Baselga, & Hyman, 2019).

Kanser tedavisini ilaç duyarlılığı sonucu olumsuz yönde etkileyen çeşitli nedenler bulunmaktadır. Bunlardan ilki ilacın farmakokinetik özellikleri, biyoyararlanımı ve tümör dokusuna ulaşımını etkileyerek ilaç duyarlılığına sebebiyet verebilecek olan ilacın yetersiz emilimi ve ulaşımı ile özetlenebilen farmakokinetik dirençtir (Longley, & Johnston, 2005). Özellikle tümör mikro çevresinde gelişen normal olmayan damar yapısı ve artan hücresel yoğunluk, ilaç dağılımını kısıtlayarak tedavi etkinliğini düşürebilir (Jain, 2005). İkinci neden effluks pompası aktivasyonu sebebiyle hücre dışına ilaç çıkışının artmasıdır. Kanser hücreleri, effluks pompası aktivasyonu sonucu ATP bağlayan kaset taşıyıcıları (ABC transporter) gibi effluks pompalarını kullanarak ilaçları hücre dışına atabilir. Özellikle P-glikoprotein (P-gp; MDR1/ABCB1), Multidrug Resistance Protein 1 (MRP1; ABCC1) ve Breast Cancer Resistance Protein (BCRP; ABCG2) kemoterapi ilaçlarının hücre içinde birikmesini önleyerek çoklu ilaç duyarlılığına neden olmaktadır (Gottesman, 2002). Üçüncü neden ilacın hücreye girişinin engellenmesi sonucu azalan ilaç alımıdır. Hücre

zarında bulunan taşıyıcı proteinler, ilaçların hücre içine alınmasını sağlar. Örneğin, organik katyon taşıyıcıları (OCT1/SLC22A1) gibi proteinlerin ekspresyonunun azalması, tirozin kinaz inhibitörleri (örneğin, imatinib) gibi ilaçların etkinliğini düşürebilir (White ve ark., 2006). Dördüncü neden ilacın metabolizmasının değişmesidir. Bazı enzim sistemleri, kanser hücrelerinin ilaçları daha hızlı metabolize etmesine ve etkisiz hale getirmesine yol açabilir. Özellikle, sitokrom P450 enzim ailesi (CYP3A4) ve glutatyon-S-transferaz (GST) gibi enzimler, bazı kemoterapi ajanlarını metabolize ederek ilaç etkinliğini azaltmaktadır (van Schaik, 2008). Beşinci neden hedef molekülde değişim veya mutasyonlardır.

Hedefe yönelik tedaviler, spesifik proteinlere bağlanarak etki gösterir. Ancak bu proteinlerde meydana gelen mutasyonlar ilaç duyarlılığına yol açabilir. Örneğin, epidermal büyüme faktörü reseptörü (EGFR) mutasyonları, tirozin kinaz inhibitörlerine karşı direnci artırmaktadır (Suda, Rivard, Mitsudomi, & Hirsch, 2017). Benzer şekilde, BCR-ABL füzyon proteininde meydana gelen T315I mutasyonu, imatinib direncine yol açmaktadır (Shah ve ark., 2007). Altıncı neden alternatif sinyal yollarının aktivasyonudur. Kanser hücreleri, hedeflenen sinyal yolunu bypass ederek alternatif yolları aktive edebilir. Örneğin, BRAF inhibitörleri (vemurafenib) uygulandığında, MEK-ERK yolunun aktif hale gelmesiyle direnç gelişebilir (Nazarian ve ark., 2010).

Yedinci neden DNA onarım mekanizmalarının aktivasyonu olup kemoterapi ilaçları genellikle DNA hasarı oluşturarak etki gösterir. Fakat DNA onarım mekanizmalarının (örneğin, homolog rekombinasyon tamiri) aktivasyonu, kanser hücrelerinin kemoterapiye karşı direnç geliştirmesine neden olabilir. Özellikle, BRCA1/2 mutasyonlarına sahip hücrelerin PARP inhibitörlerine direnç geliştirdiği gösterilmiştir (Lord, & Ashworth, 2016). Sekizinci neden apoptoz mekanizmasının baskılanmasıdır. Kanser hücreleri, apoptoz sinyalini inhibe ederek hücre ölümünü engelleyebilir. BCL-2 ailesi proteinlerinin artışı ve p53 tümör baskılayıcı gen mutasyonları, kanser hücrelerinin apoptotik mekanizmalardan kaçmasına neden olabilirler (Hanahan, & Weinberg, 2011). Dokuzuncu neden kanser hücre plastisitesi ve fenotip değişimidir.

Epitelden mezenkimale geçiş (EMT) gibi süreçler, kanser hücrelerinin fenotip değiştirerek ilaçlara karşı direnç kazanmasına neden olabilir. EMT süreci, hücrelerin

göç yeteneğini artırarak metastatik potansiyeli yükseltir ve kemoterapötik ajanlara karşı tolerans geliştirir (Singh, & Settleman, 2010). Sonuncu neden ise tümör mikro çevresi ve immün direnç sebebiyle ilaç duyarlılığı oluşabilmesidir. Tümör mikro çevresi, kanser hücrelerinin immün kaçış mekanizmalarını kullanarak ilaçlara direnç geliştirmesine neden olabilir. Örneğin, tümör hücreleri PD-L1 ekspresyonunu artırarak immün yanıtı baskılayabilir ve immünoterapilere karşı direnç kazanabilir (Topalian ve ark., 2012). Kanserde ilaç duyarlılığı, çok yönlü biyolojik mekanizmaları içeren karmaşık bir süreçtir. Direnç mekanizmalarının anlaşılması, yeni tedavi yaklaşımlarının geliştirilmesi açısından kritik öneme sahiptir. Kombinasyon tedavileri, hedefe yönelik ilaçlarla birlikte immünoterapilerin uygulanması ve direnç mekanizmalarını aşmaya yönelik yenilikçi stratejiler, kanser tedavisinde önemli gelişmeler sağlayabilir.

Kanser hücreleri heterojendir ve kemorezistans, tümör ilerlemesi ve metastazın kök hücrelerin bir alt popülasyonuna ait Kanser Kök Hücreleri (KKH) ile ilişkilendirilendirildiğine dair sağlam kanıtlar bulunmuştur (Lawson ve ark., 2018; Baik ve ark., 2017; Ganesh, & Massagué , 2021; Yang ve ark., 2020). KKH'nin özelliklerini düzenleyerek kansere yol açan mekanizmaların anlaşılması; özellikle bu hücreleri hedeflemek ve tedavi stratejileri geliştirmek kanser için tedavi sonuçlarını ve hayatta kalma oranlarını iyileştirmeyi amaçlamak araştırmacılar için aktif araştırma alanlarıdır (Yang ve ark., 2020; Batlle, & Clevers, 2017). Bununla birlikte, KKH ile ilgili az sayıda çalışma yapılmış ve yapılan araştırmalar dayalnızca bir veya birkaç spesifik kansere odaklanmakla sınırlı kalmıştır (Song ve ark., 2022). Bu bağlamda açık erişim yayınların kullanılabilirliği ve yüksek kaliteli araştırma veri setlerine erişilebilirlik kanser araştırmaları dinamiklerini geliştirmesi açısından çok önemlidir (Vesteghem ve ark., 2020).

Onkoloji alanındaki araştırmacıların karşılaştığı başlıca sorunlar arasında invaziv tekniklerle elde edilen hasta örneklerinin sınırlamalarını, yayınlanan sonuçların tekrarlanabilirlik eksikliğini, açık erişimli makalelere duyulan ihtiyacı ve hatta daha da karmaşığı bu sonuçları destekleyen veri kümelerini bulmanın zorlukları sayılabilmektedir (Rodgers, & Collings, 2021).

Hassas onkolojinin amacı, tümörlerinin genomik profilini kullanarak hastalara bireysel olarak tedaviyi uyarlamaktır. Kanser hücre hatları gibi genetik veri kümeleri,

hassas onkolojinin çok önemli bir görevi olarak ilaç duyarlılığı tahmini için en değerli kaynaklar arasındadır. Kanser hücre hatlarının büyük panelleri için çoklu omics verilerine dayanarak ilaç duyarlılığını tahmin etmek için makine öğrenimi yöntemleri kullanılmıştır. Ancak bunun nasıl yapılacağına dair kapsamlı bir kılavuz bulunmamaktadır. Kanser karmaşık bir genetik hastalıktır. Tümörlerin doğası gereği heterojenlik nedeniyle kanserin tedavisi çok zorludur. Hassas onkoloji tümörün genomik profiline göre tedavileri uyarlamayı hedeflemektedir. Hassas onkolojinin önemli bir bileşeni olan genetik, her bir hastanın genomik profilini kullanarak en etkili tedavi seçeneklerinin bulunmasını vaat etmektedir. Ancak yine de gerek araştırılan hasta grubunun sayılı ve az sayıda halka açık veri seti olması, genetik veri kümelerine etkin erişiminin olamaması nedeniyle sınırlı klinik kullanıma sahiptir. Klinik öncesi genetik veri setlerinin veri açısından zengin doğası, makine öğrenimi yaklaşımlarının geliştirilmesi yoluyla ilaç duyarlılığını in vitro ve in vivo tahmin edebilir kılmaktadır (Adam, Rampášek, Safikhani, Smirnov, Haibe-Kains, & Goldenberg, 2020). Bu hesaplamalı yaklaşımlar basit doğrusal regresyon modelleri Lasso, Elastic Net'ten, Random Forest, kernel-based modeller, Deep Neural Networks'e dayanan oldukça non-linear modeller ve en son olarak güçlendirme öğrenme, birkaç adımlı öğrenme ve çok görevli öğrenmeye kadar uzanmaktadır. Bu yöntemler genellikle gen ifadesini girdi olarak alır ve doz-cevap eğrisinin (AAC/AUC) üstündeki/altındaki alan veya yarı maksimum inhibitör konsantrasyon (IC50), konsantrasyon canlılığı %50 oranında azaltan ilaç tahmininde bulunur (Sharifi-Noghabi ve ark., 2021).

2.4. Sağlık Alanında Makine Öğrenmesi

Makine öğrenmesi (ML), hastalık teşhisi, tedavi planlaması, kişiselleştirilmiş tıp, sağlık yönetimi gibi sağlık alanında devrim yaratacak pek çok sağlık alanında uygulanmaktadır. Büyük veri analizi, görüntü işleme ve doğal dil işleme gibi tekniklerle sağlık hizmetlerini iyileştirmek ve maliyetleri azaltmak mümkündür.

Yapay zeka son on yılda hem akademik anlamda hem de dışında popüler bir konu haline geldi; teknoloji ve teknoloji dışı dergilerdeki çok sayıda makale makine öğrenimi, derin öğrenme ve yapay zeka konularını ele aldı (Brown ve ark., 2018;

Gulshan ve ark., 2016; Coyner ve ark., 2019; Rajpurkar ve ark., 2017; Jones, Golan, Hanna, & Ramachandran, 2018; De Fauw ve ark., 2018).

Yapay zeka normalde insanlar tarafından gerçekleştirilen entelektüel görevlerin otomasyonuna odaklanmış bir alan olup makine öğrenmesi ve derin öğrenme bu hedefe ulaşmanın uygun yöntemleridir ve yapay zeka kapsamı dahilindedir. Ancak, yapay zeka herhangi bir "öğrenme" biçimini içermeyen yaklaşımları içerir. Örneğin, sembolik yapay zeka olarak bilinen alt alan, belirli bir ilgi alanındaki her olası senaryo için kuralların sabit kodlanmasına yani açıkça yazılmasına odaklanır. İnsanlar tarafından yazılan bu kurallar, tamamlanması gereken belirli konu ve göreve ilişkin önceden öğrenilmiş bilgiden gelir. Örneğin, bir kişi bir ofisin oda sıcaklığını uygun olarak belirlemek için bir algoritma oluşturacak olsaydı, o kişi muhtemelen insanların çalışmak için hangi sıcaklıklarda rahat olduğunu belirler ve sıcaklıkları belirli bir üst değerin üzerine çıktığında odayı soğutacak ve alt değerin altına düştüğünde ısıtacak şekilde programlamasını yapardı. Sembolik yapay zeka, açıkça tanımlanmış problemleri çözmede başarılı olsa da, konuşma tanıma veya görüntü sınıflandırması gibi daha yüksek seviyeli desen tanıma gerektiren görevlerde sıklıkla başarısız olur. Bu görevler makine öğrenmesi ve derin öğrenme yöntemlerinin iyi performans gösterdiği daha karmaşık görevlerdir (Choi, Coyner, Kalpathy-Cramer, Chiang, & Campbell, 2020).

Makine öğrenmesi, istatistiksel yöntemler ve algoritmalar kullanarak verilerden öğrenme sürecidir. Sağlık alanında kullanılan temel makine öğrenmesi teknikleri şu şekilde sınıflandırılabilir;

- ✓ Denetimli Öğrenme: Etiketli veri kullanılarak model eğitilir. Örneğin, hastalık teşhisi için lojistik regresyon ve karar ağaçları kullanılabilir.
- ✓ Denetimsiz Öğrenme: Etiketlenmemiş veri setlerinden örüntüler çıkarılır. Kümeleme algoritmaları ile hastalık sınıflandırması yapılabilir.
- ✓ Takviyeli Öğrenme: Deneme-yanılma yöntemi ile en iyi kararları öğrenen sistemler geliştirilir. Robotik cerrahi ve ilaç keşfi alanlarında kullanılmaktadır (Rauschert, Raubenheimer, Melton, & Huang, 2020). Özellikle genetikte yararlanılan makine öğrenmesi yöntemleri Tablo 2' de gösterilmektedir.

Tablo 2. Genetikte kullanılan başlıca makine öğrenmesi yöntemleri

Makine Öğrenmesi Yöntemleri ve Biyomedikal Kullanım Alanları

Yöntem	Kullanım Alanı
Destek Vektör Makineleri (SVM)	Biyobelirteç keşfi ve ilaç yanıtının tahmin edilmesi
Rastgele Ormanlar (Random Forest)	Genetik risk faktörlerinin belirlenmesi
Derin Öğrenme (CNN, RNN, Transformer)	Genom dizilim analizi ve advers ilaç reaksiyonlarının tahmini
Kümeleme (K-Means, DBSCAN)	Hasta alt gruplarının belirlenmesi
Takviyeli Öğrenme (Reinforcement Learning)	Kişiselleştirilmiş tedavi stratejilerinin oluşturulması

Makine öğrenmesi, büyük ölçekli genomik ve klinik verileri analiz ederek ilaç yanıtlarını tahmin etme, biyobelirteç keşfetme ve kişiselleştirilmiş tedavi geliştirme süreçlerinde önemli bir rol oynamaktadır. Sağlık hizmetlerini daha verimli ve etkili hale getirme potansiyeline sahiptir. Ancak, bu teknolojilerin güvenilirliği ve etik kullanımı üzerine daha fazla çalışma yapılması gerekmektedir.

2.5. İlaç Duyarlılığının Tahmin Edilmesinde Makine Öğrenmesi

Çok yeni makine öğrenimi teknikleri; öneri sistemleri, sıralama yöntemleri, üretken modeller, özellik analizi, ağ modelleme, topluluk modelleri ve derin öğrenme yaklaşımları ile genomik özellikler araştırılmaktadır. Yakın zamanda ilaç yanıtı tahmini, önemli örtüşme ve verilere özel vurgu ile entegrasyon, özellik seçimi, deneysel karşılaştırma, makine öğrenme yöntemleri, sistematik kıyaslama, kombinasyon tedavisi, derin öğrenme sonuçları ve meta-inceleme konusunda bir çok inceleme makalesi yayınlanmıştır.

İlaç duyarlılığının tahmin edilmesinde makine öğrenmesi, hastaların tedaviye yanıtlarını öngörmek, yeni antibiyotikler geliştirmek ve kişiselleştirilmiş tıbbı ilerletmek için giderek daha fazla kullanılan bir yaklaşımdır. Bu süreçte, çeşitli makine öğrenmesi algoritmaları kullanılarak, hastalığa neden olan mikroorganizmaların belirli ilaçlara karşı geliştirdiği direnç düzeyi tahmin edilebilir. İlaç duyarlılığının tahmin edilmesinde kullanılan makine öğrenmesi tekniklerinden belli başlı olanları lojistik regresyon (Logistic Regression), karar ağaçları (Decision Trees) ve rastgele ormanlar (Random Forests), destek vektör makineleri (Support Vector Machines - SVM) ve derin öğrenmedir (Deep Learning).

2.6. Farmakogenomiğin Ortaya Çıkışı

Farmakogenomik, bireylerin genetik farklılıklarının ilaçlara verdikleri yanıtları nasıl şekillendirdiğini inceleyen bir bilim dalıdır. Yapay zeka alanında kaydedilen hızlı gelişmeler neticesinde genetik verinin analizi, sağlık alanlarına katkısı artmış olup gelecek açısından da ümit vaat etmektedir.

Makine öğrenmesinden genetik biliminde pek çok açıdan yararlanılmaktadır. İlaç yanıtlarının tahmin edilmesi bu alanlardan biridir. İnsan DNA'sındaki genetik varyasyonlar, bireylerin ilaçlara duyarlılıkları veya dirençlerinde belirleyici rol oynayabilir. Makine öğrenmesi teknikleri, genetik verileri analiz ederek hastaların ilaçlara nasıl tepki vereceğini tahmin etmek için kullanılmaktadır. Feng ve arkadaşlarının 2022 yılında yaptığı çalışmada, ilaç yanıtı tahmininde AGMI isimli yeni bir model geliştirilmiştir. AGMI (Attention-Guided Multi-omics Integration), her hücre hattı için bir Multi-edge Graph (MeG) oluşturur ve multi-omik özellikleri toplayarak ilaç yanıtını tahmin etmek için Graph edge-aware Network (GeNet) yapısını kullanır. Bu yaklaşım, bütün genomu kullanarak ilaç yanıtı tahmininde gen kısıtlamalı çoklu omik entegrasyonunu keşfetmek için bir alternatif sunabilir.

Genomics of Drug Sensitivity in Cancer (GDSC) ve Cancer Cell Line Encyclopedia (CCLE)'deki veri setlerini kullanarak kanser hastalarının genetik profillerini temel alan ve ilaç yanıtlarını tahmin etmeye çalışan derin öğrenme modelleri geliştirmek yeni bir yaklaşım sunmaktadır. Makine öğrenmesi, ilaç duyarlılığı veya ilaçlara karşı gelişen dirençle ilişkili genetik biyobelirteçleri belirlemek için de kullanılır. Şentürk ve arkadaşlarının 2021 yılında yaptığı çalışmada, BRCA1 ve BRCA2 gen varyasyonlarının meme kanseri riskini değerlendirmek amacıyla yapay zeka modelleri geliştirilmiştir. Toplamda 268 meme kanseri hastasının verileri kullanılarak, 16 farklı risk faktörü dahil edilmiştir. BRCA1 (61 varyant), BRCA2 (128 varyant) ve her iki gene ait 11 varyantın verileri, MATLAB platformunda Mamdani Fuzzy Mantık Sistemi ve Feed-Forward Sinir Ağı Yöntemi ile modellenmiştir. Geliştirilen modeller, genetik verilerin kullanımının, kişiselleştirilmiş tıp uygulamaları için umut verici sonuçlar sunabileceğini ifade etmiştir.

Gen-ilaç ve ilaç-ilaç etkileşimlerinin analizinde de genetikte makine

öğrenmesinden fayda sağlanmaktadır. Bazı ilaçlar genetik faktörlere bağlı olarak farklı yan etkilere neden olabilir. Advers ilaç reaksiyonlarının tahmininde de yapay zeka etkin rol oynamaktadır. Yanlış ilaç kullanımı ciddi sağlık sorunlarına yol açabilir. Doğal Dil İşleme (NLP) ve derin öğrenme teknikleri, elektronik sağlık kayıtları (EHR) ve farmakovijilans verilerini analiz ederek advers ilaç reaksiyonlarını öngörmek için kullanılmaktadır. Yapay zeka uygulamalarından kişiye özel ilaç dozajı ayarlanması konusunda da destek alınması hastaların tedavisinde önem arzetmektedir. Bireylerin metabolizma hızları farklı olduğundan, ilaç dozlarının kişiye özel olarak ayarlanması gerekmektedir. Makine öğrenmesi modelleri, genetik varyasyonlar ve klinik parametreleri dikkate alarak en uygun ilaç dozlarını belirleyebilir. Johnson ve arkadaşları 2018 yılında yaptıkları çalışmada warfarin dozu tahmini için makine öğrenmesi modelini uygulamış ve VKORC1 ve CYP2C9 genlerindeki varyasyonlar analiz edilerek bireye özgü dozaj önerileri geliştirilmiştir.

Makine öğrenmesi modellerinin yüksek doğrulukla çalışabilmesi için büyük ve kaliteli veri setlerine ihtiyaç vardır. Ancak, genetik veri kaynakları genellikle sınırlıdır. Klinik karar destek sistemlerinde kullanılacak modellerin açıklanabilir ve şeffaf olması etik açıdan da açıklanabilir olmalıdır ve bu konuda geniş çaplı düzenlemelerin yapılması önemlidir.

2.7. Mutasyon Verisi ve İlaç Yanıtı Tahminindeki Önemi

Genomik seviyede meydana gelen değişikliklerle bazı ilaçların oluşturabileceği olası yanıtlar arasında bir korelasyon bulunmaktadır. Bu düşünceyle, tedavide kullanılan ilaçlara karşı gelişebilecek yanıtın tahmininde hastaya ait genetik varyasyonların da yararlı olabileceği ile ilgili eğilim gelişmiştir (Kang, & Hong, 2011; Zhou ve ark. 2021). Böylece, hastaya özel olacak şekilde bireyselleştirilmiş ilaç, doz belirlenmesini kolaylaştıracak ve yan etki riski azaltılabilecektir (Kang, & Hong, 2011). Hücre bazında bir ilacın etkili olabilmesi için hücre içinde bazı süreçlerden geçip hedeflendiği yapılara bağlanarak işlevini yerine getirmesi gerekmektedir. Hücredeki ilacı metabolize eden enzimlerin, ilaç taşıyıcı proteinlerin veya hedef proteinlerin yapısında değişiklik olması durumunda ilaç yanıtı değişebilecektir. İlgili proteinleri kodlayan genetik sekanslarda oluşacak değişimler

üründeki yapı değişikliklerine veya proteinin sentezlenmesinin engellenmesine neden olabilmektedir (Kang, & Hong, 2011; Zhou ve ark., 2021; Naulaerts, Dang, & Ballester, 2014). Tek nükleotid polimorfizmleri (single nucleotide polymorphism, SNP) somatik hücrelerde oluşan genetik varyasyonların ana sebeplerindendir. SNP'ler, tek baz çifti değişimi mutasyonu olarak DNA sekansındaki her 1000 baz çifti içinde ortalama frekansı 1 SNP olacak şekilde meydana gelirler. Bilinen SNP'lerin % 1 kadarı, protein kodlayıcı sekansta değişimler oluşturur. Bu değişimle, protein yapısına eklenen farklı bir amino asit ile proteinin fonksiyonunda varyasyonlar (yüksek aktivite veya inaktivasyon gibi) oluşabilmektedir. Ayrıca DNA sekansındaki değişimler amino asit ifade eden kodonun bir durdurucu kodona dönüşmesine neden olarak proteinin normalden hızlı şekilde bozunmasına veya inaktif olmasına neden olabilirler. Sessiz SNP'ler ise protein kodlayan sekanslardaki mutasyon değişimine proteinde işlevsel olarak herhangi bir etki oluşturmamaktadır (Kang, & Hong, 2011).

Literatürde mutasyon verilerinin ilaç yanıtı tahmininde olumlu etkisini gösteren çeşitli çalışmalar bulunmaktadır. Bu çalışmalar, genetik mutasyonların belirli ilaçlara karşı duyarlılığı veya direnci nasıl etkilediğini ortaya koyarak, tedavi stratejilerinin kişiselleştirilmesine önemli ölçüde katkıda bulunabilmektedirler. Bu çalışmada, mutasyon verisi dahil olmak üzere tüm genomik profillerin entegre edilmesiyle, derin öğrenme modelleri kullanılarak tümörlerin ilaç yanıtı tahmin edilmiştir. Mutasyon verisi, tahmin performansını artıran en önemli biyolojik bileşenlerden biri olarak gösterilmiştir. Yapılan bir çalışmada derin öğrenme modelleri kullanılarak tümörlerin genetik profilleri analiz edilmiş ve bu modellerin, mutasyon ve gen ekspresyon verilerini entegre ederek ilaç yanıtını tahmin etmede başarılı olduğu gösterilmiştir. Tümör ilaç yanıtlarını tahmin etmek için gen ekspresyonu**,** mutasyon**,** ve kopya sayı varyasyonları (CNV) gibi çoklu genomik veriler kullanılarak derin sinir ağı (deep neural network) modelleri geliştirilmiştir. Mutasyon verisinin modele dahil edilmesi, tahmin doğruluğunu anlamlı şekilde artırmıştır. EGFR ve BRAF gibi onkogenlerdeki mutasyonlar, bu genlerin hedef alındığı ilaçlara duyarlılık tahminlerinde biyobelirteç olarak kullanılmıştır. Derin öğrenme modelinin performansı, klasik makine öğrenmesi yöntemlerinden (Random Forest, SVM) daha iyi bulunmuştur. Mutasyon verileri, hücrelerin ilaçlara karşı nasıl yanıt vereceğini belirlemede güçlü belirteçler

olarak kabul edilmiş ve modelin genelleştirilebilirliğini artırmıştır (Chiu ve ark., 2019).

Bir başka çalışmada mutasyonların nicel skorlarla temsil edilmesinin, ilaç yanıtı tahmin modellerinin doğruluğunu artırdığı gösterilmiş ve BRAF V600 mutasyonları gibi spesifik genetik değişikliklerin belirli tedavilere duyarlılığın daha doğru tahmin edilmesi sağlanmıştır. Çalışma, kanser ilaç yanıtı tahmininde mutasyonları sadece "var/yok (1/0)" şeklinde değil, fonksiyonel skorlarla (quantitative mutation scores) temsil etmenin model performansını artırdığını göstermektedir. Çalışmada kullanılan skorlayıcılar; CHASMplus, VEST4, CADD – olup bu algoritmalar, bir mutasyonun gen fonksiyonunu ne kadar etkilediğini nicel olarak tahmin eder. Örnek olarak, BRAF-V600 mutasyonuna karşı dabrafenib yanıtı daha doğru şekilde tahmin edilmiştir. Yöntem ayrıca EGFR, ERBB2, PIK3CA gibi genler için de genellenebilir başarı göstermiştir. Mutasyonları skorlayarak kullanmak, sadece mevcut biyobelirteçlere değil, aynı zamanda yeni ilaç-biyobelirteç ilişkilerinin keşfine de olanak tanıyabilecektir (Wall, & Ideker, 2024).

Diğer bir çalışmada akut miyeloid lösemi (AML) hastalarından alınan binden fazla örnekte, ilaç yanıtı tahmini için mutasyon örüntüleri analiz edilmiştir. Mutasyon verisi**,** özellikle FLT3, NRAS, KRAS, TP53 gibi genlerde, belirli ilaçlara duyarlılığı yüksek doğrulukla tahmin etmiştir. RAS mutasyonları, özellikle MEK inhibitörleri (örneğin trametinib) gibi hedefe yönelik ajanlara karşı duyarlılığın güçlü bir göstergesi olarak gösterilmiştir. Kombine modellerde mutasyon ve ekspresyon verisi birlikte kullanıldığında, model performansı daha da yükselmiştir. Çalışma, klinik olarak uygulanabilir mutasyon imzalarının (mutation signatures) tedaviye yön vermede etkili olabileceğini göstermektedir. Ayrıca ilaç yanıtı tahmini açısından biyobelirteç keşfinde mutasyon verisinin değerini vurgulamıştır (Qin ve ark., 2024). Bu çalışmalar, mutasyon verilerinin gelecekte ilaç yanıtı tahmininde kritik bir rol oynayacağını ve kişiselleştirilmiş tedavi yaklaşımlarının geliştirilmesine katkı sağlayabileceğini göstermektedir.

2.8. Kopya Sayısı Değişimi Verisi ve İlaç Yanıtı Tahminindeki Önemi

Kopya Sayısı Değişimleri (Copy Number Variations - CNV), genomda belirli

DNA segmentlerinin sayısındaki farklılıkları ifade eder ve kanser gelişiminde önemli bir rol oynar. CNV'ler, onkogenlerin amplifikasyonu veya tümör baskılayıcı genlerin delesyonu yoluyla tümör oluşumunu tetikleyebilir. Ayrıca, CNV'lerin varlığı, kanser tedavisinde kullanılan ilaçlara verilen yanıtı etkileyebilir ve bu nedenle, kanser tedavisinin kişiselleştirilmesinde önemli bir faktör olarak değerlendirilir.

Yapılan bir çalışmada, insan genomunda yer alan yapısal varyasyonların küresel olarak nasıl dağıldığı incelenmiştir. Yapısal varyasyonlar, genetik materyalin eklenmesi, çıkarılması veya duplikasyonu gibi büyük değişimleri kapsamaktadır. Çalışmada, dünya çapında farklı popülasyonlardaki insan genetik yapısındaki farklılıkları incelenerek, CNV yaygınlığı; CNV'lerin genomda ne kadar sık görüldüğü ve bu değişikliklerin genetik hastalıklar, kanser gibi durumlardaki rolü araştırılmıştır. CNV'lerin, bireyler arasındaki genetik çeşitliliğin büyük bir kısmını oluşturduğu ve bunun tedavi anlamında fark yaratabileceği çalışmada incelenmiştir (Conrad ve ark., 2010). Redon ve arkadaşlarının 2006 yılında yaptığı araştırmada, insan genomunda kopya sayısı değişimlerinin dünya genelindeki yaygınlığını incelenmiştir. Araştırmacılar, farklı bireylerin genomlarında bulunan CNV'leri analiz ederek, bu değişikliklerin insan sağlığına ve genetik hastalıkların gelişimine etkilerini incelemiş ve kopya sayısı değişimlerinin, genomdaki bazı genetik hastalıkların ve kanser türlerinin nedenlerini anlamamıza yardımcı olabileceğini öne sürmüşlerdir. CNV'lerin bireysel genetik farklılıklar oluşturduğunu ve kişiye özel tedavi stratejilerinin geliştirilmesine katkı sağlayabileceği bir başka çalışmada belirtilmiştir (Meyerson, Gabriel, & Getz, 2010).

Bir diğer makalede, kanserin moleküler temellerini inceleyen The Cancer Genome Atlas (TCGA) projesi ele alınmıştır. TCGA, kanser türlerinin genomik haritalarını çıkararak, her kanser türünde görülen kopya sayısı değişimlerini araştırmaktadır. Çalışmada, HER2 gibi onkogenlerin amplifikasyonunun kanser gelişimiyle nasıl ilişkilendiği ve tedaviye olan etkileri vurgulanmıştır. Özellikle HER2 amplifikasyonunun meme kanseri tedavisinde önemli bir biyobelirteç olduğu ve bu bilgilerin tedavi stratejilerini şekillendireceği makalede ifade edilmiştir. Ayrıca, CNV'lerin kanser tedavisindeki rolü ve tedaviye direnç gibi durumlar üzerindeki etkisi de makalede tartışılmıştır. Yapılan başka bir çalışmada, somatik kopya sayısı değişimlerinin kanser türleri arasındaki yaygınlığı araştırılmış ve kanser

hücrelerinde görülen kopya sayısı değişikliklerinin, tümörlerin evriminde nasıl bir rol oynadığı tartışılmıştır. Çalışmada, CNV'lerin kanserin gelişiminde ve tedaviye yanıtında önemli etkiler yaratabileceği ifade edilmiştir. Somatik CNV'ler, kanserin moleküler biyolojisini anlamada kritik bir bileşen olarak kabul edilmiş ve bu değişikliklerin hangi genlerde yoğunlaştığına dair detaylar sunulmuştur (Beroukhim ve ark., 2010).

Bu çalışmalar, CNV'nin kanser, genetik hastalıklar ve evrimsel biyoloji gibi farklı alanlarda nasıl önemli bir rol oynadığını ve bunun klinik uygulamalara nasıl dönüştürülebileceğini anlamamıza yardımcı olabilecektir. CNV' lerin bireysel tedavi stratejilerinde etkili birer biyobelirteç olabileceği ve bu değişimlerin genetik hastalıkların anlaşılmasında kullanılabileceği fikri birçok araştırmaya ilham olmaktadır.

2.9. Moleküler Biyoloji, Genomik ve Hesaplama Teknolojilerindeki İlerlemeler

Moleküler biyoloji, genomik ve hesaplama teknolojilerinde kaydedilen hızlı ilerlemeler, farmakolojik araştırmalarda önemli gelişmelerin gerçekleşmesine sağlam bir temel oluşturmuştur. Bu alanların bütünleşmesi ise özellikle ilaç keşfi ve ilaçların yeniden konumlandırılması süreçlerinde ilerlemeleri hızlandırmıştır (Shaker, Tran, Jung, & Na, 2021; Shaker, Ahmad, Lee, Jung, & Na, D, 2021; Zeng ve ark., 2022).

Son yıllarda yapılan ilaç araştırmalarında, yeni ilaçların keşfi ve mevcut ilaçların yeniden konumlandırılması ön plana çıkmıştır (Wan ve ark., 2025). Yeniden konumlandırma stratejisi, mevcut ilaçların tanımlı endikasyonlarının dışında yeni tedavi alanlarında kullanılabilmesine olanak sağlamaktadır (Pushpakom ve ark., 2019; Wan ve ark., 2025). Bu yöntem sayesinde, ilaç geliştirme sürecinde erken evrelerde yer alan keşif, preklinik deneyler ve üretim hazırlıkları gibi aşamaların büyük ölçüde aşılması mümkün olmakta; böylece süreç daha düşük maliyet, daha az zaman ve daha az emek gerektirmektedir (Ashburn & Thor, 2004). Nitekim, geçmişte farklı hastalıklar için kullanılan minoksidil, talidomid ve sildenafil günümüzde yeniden konumlandırılarak farklı hastalıkların tedavisinde kullanılmaktadır (Iyer ve ark., 1971; Convit ve ark., 1970; Wan ve ark., 2025).

İlaç yeniden konumlandırma süreci incelendiğinde, ilaç duyarlılığı kavramının

ön plana çıktığı görülmektedir. İlaç duyarlılığı, çoğunlukla kanser hücrelerinin çeşitli hücresel mekanizmalar aracılığıyla ilacın etkisini engellemesi ya da yanıtsız kalması sonucunda, ilacın beklenen terapötik yararı sağlayamaması şeklinde tanımlanmaktadır (Vasan ve ark., 2019; Mao ve ark., 2025). Bu durum, ilaç tedavilerinin başarısız olmasına yol açabilmektedir. Dolayısıyla, ilaçların etkinliği ve tedaviye verilen yanıt, ilaç molekülleri ile biyolojik sistem arasındaki karmaşık etkileşimlerden ötürü hastadan hastaya değişiklik gösterebilmektedir (Wang, McLeod, & Weinshilboum, 2011; Ha, Park, & Jo, 2025).

2.10. Kişiselleştirilmiş Tıp

Genomik ile bilgisayar biliminin kesişiminden doğan çalışmalar, ilaçların hastalar üzerindeki duyarlılık etkilerinin tahmin edilmesini mümkün kılmakta ve bu bilgiler ilaç yeniden konumlandırma stratejilerinde kullanılabilmektedir. İlaç yanıtının öngörülmesine yönelik araştırmalar, özellikle kanser tedavisinde kişiselleştirilmiş tıp uygulamalarının geliştirilmesinde büyük önem taşıyan, görece yeni bir bilimsel alan olarak değerlendirilmektedir (Meng, Xu, Xiao, Gao, & Yu, 2025; Ha, Park, & Jo, 2025; Mao ve ark., 2025; Carli ve ark., 2025). Yüksek kapasiteli teknolojilerin hızla ilerlemesi ve genetik verilerin sürekli artması sayesinde, tedaviye uygun ilaç moleküllerinin belirlenmesi çok daha kolay hale gelmektedir.

2.11. Onkolojide Sürücü Mutasyonlar ve Genetik Yaklaşımlar

Onkolojide genomik verilerden yararlanılarak, tümörlerdeki sürücü mutasyonların saptanmasına yönelik hedeflenmiş gen dizileme panelleri geliştirilebilmektedir. Hücre fonksiyonlarını bozan mutasyonlar, kanserin en temel biyolojik özelliklerinden biri olarak kabul edilmektedir (Najgebauer, Perron, & Iorio, 2021; Stratton, Campbell, & Futreal, 2009). Sürücü mutasyonlar, normal hücrelerin malign özellikler kazanarak kanser hücrelerine dönüşme süreçlerinde rol oynayan temel değişiklikleri açıklamaktadır (Najgebauer ve ark., 2021).

Wang, Fang, & Chen (2016) kanserle ilişkili genomik değişiklikleri ilaçların

kimyasal özelliklerine dayalı genetik verilerle bütünleştirmiştir. Bu çalışma, terapötik yanıt tahminlerinin daha doğru yapılabilmesi için çoklu genetik veri kaynaklarını bir araya getiren kapsamlı yaklaşımlara duyulan ihtiyacı ön plana çıkarmıştır.

2.12. İlaç-İlaç Etkileşimleri ve Hesaplamalı Yaklaşımlar

İlaç-ilaç etkileşimleri, tedavilerin güvenliği ve etkinliği açısından önemli bir sorun oluşturmaktadır. Bu bağlamda, DrugBank veri seti kullanılarak geliştirilen stacking ensemble modeli, %99 doğruluk sağlayarak klasik yöntemlerin üzerine çıkmış ve derin öğrenme yaklaşımlarıyla daha da geliştirilebileceğini göstermiştir (Abbas ve ark., 2023). Buna ek olarak, yakın dönemde gerçekleştirilen bir çalışmada Glioblastoma için ilaç etkinliği öneri sistemi geliştirilmiştir. GDSC veri seti kullanılarak gen ekspresyonu, ilaç SMILES ve IC50 parametreleri bir araya getirilmiş; CNN modelinin, ANN'e kıyasla daha yüksek performans gösterdiği ortaya konmuştur (Naveed ve ark., 2025).

Mutasyonlar ve DNA metilasyonu gibi genetik özelliklerle ilaçlar arasındaki ilişkiyi anlamak için kanser hücre hatları önemli bir model olarak kullanılmaktadır. Iorio ve arkadaşları (2016), hücre hatlarının genetik yapısını ve ilaçlara karşı verdikleri yanıtları sistematik biçimde incelemiş ve bu modellerin, gerçek tümörlerdeki genetik değişimlerin büyük bölümünü yansıttığını göstermiştir. Bulgular, bazı genetik değişiklik kombinasyonlarının hücreleri belirli kimyasal bileşiklere karşı daha hassas hale getirdiğini, ayrıca makine öğrenmesi yöntemlerinin ilaç yanıtı tahmininde farklı moleküler veri tiplerinin rolünü ortaya koyduğunu göstermektedir.

Araştırmacılar son yıllarda CCLE, GDSC ve CTRP gibi açık veri kaynakları sayesinde geniş ölçekte veri erişimine kavuşmuştur. Özellikle CCLE (Barretina ve ark., 2012) ve CGP (Garnett ve ark., 2012) çalışmaları, klinik açıdan önemli hücre hatlarını ve bunların kanser ilaçlarına ilişkin farmakolojik profillerini analiz ederek potansiyel biyobelirteçlerin sistematik biçimde belirlenmesini sağlamıştır. Buna paralel olarak, Zhang ve arkadaşları (2015) ile Yang ve arkadaşları (2013), CCLE ve CGP veri tabanlarını kullanarak hücre hatlarının ilaç yanıtlarını değerlendirmiş ve benzer gen ekspresyon örüntülerine sahip hücre hatlarının aynı ilaca benzer tepkiler

verdiğini göstermiştir. Ayrıca, yeni hücre hatlarının mevcut ilaçlara ya da mevcut hücre hatlarının yeni ilaçlara tepkisinin öngörülebileceğini ortaya koymuştur.

2.13. Veri Entegrasyonu ve Makine Öğrenmesi Modelleri

Deneysel süreçlere gerek kalmadan ilaç yanıtının tahmin edilebilmesi için yapay zekâ yöntemlerinden yararlanılmaktadır. Long ve arkadaşları (2025), kanser tedavisinde ilaç duyarlılığının tahminine yönelik yapay zekâ uygulamalarını kapsamlı biçimde ele almıştır. Öte yandan, Carli ve arkadaşları (2025), GDSC ve PRISM veri setlerini kullanarak ilaçların etki mekanizmalarını öğrenebilen ve bu bilgiyi hasta örneklerine genelleyebilen bir makine öğrenmesi modeli geliştirmiştir.

Makine öğrenmesi, verilerden anlamlı örüntüler çıkarmaya olanak tanıyan bir yöntemdir (Hastie, Tibshirani, & Friedman, 2009). Bu alandaki araştırmaların kökeni 1950'lere kadar uzanmakta olup (Alpaydın, 2020), günümüzde bir bağımlı değişken ve çok sayıda bağımsız değişken içeren tablo verilerinde yüksek doğruluk elde etmeye imkân tanımaktadır. Dolayısıyla, açık kaynaklı veri kümelerinden yararlanarak ilaç etkileşimlerini öngörebilen modellerin geliştirilmesi mümkündür.

Ha ve arkadaşları (2025), GDSC veri seti üzerinde 13 farklı regresyon modelini test etmiş ve en iyi performansın destek vektör regresyonu (SVR) ile sağlandığını rapor etmiştir. Çalışma ayrıca, hormonla ilişkili yolakların bilgi açısından, kopya sayısı değişimleri (CNV) ve mutasyonlardan daha açıklayıcı olduğunu göstermiştir. Elde edilen bulgular, özellik seçiminin kritik önemini ortaya koymaktadır.

GDSC verileri kullanılarak yapılan bir diğer çalışmada, ilaç duyarlılığı ile mutasyon yükü arasındaki bağlantı araştırılmış ve Gradient Boosting algoritmasına dayalı bir model önerilmiştir (Mergen, & Özcan, 2025). Her ne kadar sonuçlar umut verici olsa da, yalnızca tek bir veri kaynağına dayanılması, modelin genellenebilirliği açısından sınırlılık oluşturmaktadır.

Gradient Boosting algoritması, fonksiyonel uzayda "boosting" yaklaşımını temel almakta ve genellikle basit karar ağaçlarından oluşan zayıf öğrenicilerin bir araya getirilmesiyle çalışmaktadır (Friedman, 1999). Karar ağaçlarının zayıf öğrenici olarak kullanılmasıyla ortaya çıkan bu yaklaşım gradient-boosted trees olarak adlandırılmaktadır. Son dönemde geliştirilen LightGBM, XGBoost ve CatBoost gibi

algoritmalar, hem sınıflandırma hem de regresyon problemlerinde yüksek başarı sağlamaktadır (Ke ve ark., 2017; Chen, & Guestrin, 2016; Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin, 2018).

Bunun yanı sıra, derin öğrenme ve transfer öğrenme yöntemleri de ilaç duyarlılığı tahmininde dikkat çekici ilerlemeler sağlamaktadır. Meng ve arkadaşları (2025), farklı veri kümeleri arasındaki dağılımsal farklılıkları azaltmak için transfer öğrenme tabanlı bir yöntem geliştirmiş ve modeli GDSC ile CCLE gibi büyük ölçekli veriler üzerinde eğitmiştir. Öte yandan, Bueschbell ve arkadaşları (2022), ağ biyolojisi ile yapay zekânın rolünü incelemiş; genomik, epigenetik, transkriptomik, proteomik, metabolomik ve genetik gibi omik verilerin bütüncül biçimde işlenmesi gerektiğini vurgulamıştır. Sonuç olarak, moleküler düzeyde geliştirilen yapay zekâ temelli araçlar, kanser tanı ve tedavisinde geleneksel yöntemlerin etkinliğini artırma potansiyeline sahiptir. Ancak, kanserin ortaya çıkışı, ilerleyişi ve ilaçlara yanıt mekanizmaları karmaşık bir yapıya sahiptir; mevcut veri kaynakları bu süreci tam anlamıyla açıklamak için henüz yeterli değildir.

Moleküler ölçekte geliştirilen yapay zekâ tabanlı yöntemler, kanser tanı ve tedavisinde kullanılan klasik yaklaşımların etkinlik ve verimliliğini artırma potansiyeline sahiptir. Bununla birlikte, kanserin biyolojik oluşum süreci, ilerleyişi ve ilaçlara duyarlılığı oldukça karmaşık mekanizmalara dayandığından, mevcut veri kaynakları bu süreci tam olarak aydınlatmak için hâlâ yetersizdir. Mergen, Çoban, Özkan, Başaran, & Özcan'ın 2026 yılında yaptığı bir çalışmada geniş kapsamda, tüm kanser hücre hatlarına ait ilaç duyarlılığı verileri, somatik mutasyon verileri ve gen–ilaç etkileşim bilgilerini içeren farklı veri tabanları entegre edilmiş ve bu konuda literatüre katkı sunulmuştur.

Son yıllarda yapay zekâ ve makine öğrenmesi, ilaç keşfinde devrim niteliğinde değişimlere yol açmış; ancak bu teknolojilerin giderek karmaşık hale gelmesi, yorumlanabilirlik ve şeffaflık ihtiyacını daha da artırmıştır (Alizadehsani ve ark., 2024). İlaç geliştirme alanında kuantum bilişim de öne çıkan bir teknoloji haline gelmiştir. Klasik bilgisayarlarla çözülmesi zor olan farmasötik problemlerin, disiplinler arası yaklaşımlar çerçevesinde geliştirilen kuantum algoritmaları ile çözülebileceği öngörülmektedir (Jain, Chouzenoux, Kumar, & Majumdar, 2023; Kumar, Yadav, Mukherjee, Hassija, & Guizani, 2024).

İlaç duyarlılığı tahmininde doğruluğu artırmak için farklı veri kaynaklarının bir araya getirilmesi, etkin bir veri zenginleştirme stratejisi olarak öne çıkmaktadır. Bu çalışma kapsamında, gen mutasyonları, gen–ilaç etkileşimleri ve ilaç duyarlılığına ilişkin veriler birleştirilerek yeni biyolojik örüntüler ortaya çıkarılması hedeflenmiştir. Bunun için GDSC veri setinden ilaç duyarlılığı, COSMIC'ten mutasyon profilleri ve CNA, ayrıca DGIdb'den gen–ilaç etkileşimleri alınmış ve uygun kodlayıcılarla entegre edilmiştir. Dört farklı veri setinin entegrasyonundan doğan sorunları çözmek üzere özel veri tamamlama ve filtreleme teknikleri kullanılmıştır. Ardından, entegre veri seti sayısallaştırılmış ve en anlamlı özellikler PSO (particle swarm optimization) yöntemi ile seçilmiştir. Böylelikle, sonraki makine öğrenmesi çalışmalarında daha yüksek performans sergileyebilecek ve yorumlanabilirliği artmış bir veri seti elde edilmiştir.

Bu zenginleştirilmiş veriler üzerinde hiyerarşik kümeleme analizi yapılarak ilaçlar, ilişkili oldukları gen yanıt profillerine göre gruplandırılmıştır. Ayrıca, farklı ilaçlara verilen gen yanıtlarının etkinliği değerlendirilmiştir. Son aşamada, işlenen veri seti XGBoost, CatBoost, LightGBM, Random Forest ve Extra Trees gibi makine öğrenmesi modellerini eğitmek için kullanılmıştır. Tüm bu bulgular ışığında, entegrasyon ve optimizasyon temelli yaklaşımın, ilaç duyarlılığı tahmininde daha güvenilir ve genellenebilir modeller ortaya çıkaracağı düşünülmektedir.

3. GEREÇ VE YÖNTEM

Tez çalışmam iki aşamadan oluşmaktadır. Çalışmanın ilk aşamasında makine öğrenmesi yöntemleri ile verinin manipülasyonu, öğrenme sonucunda tahmin ve mutasyonel yüke bağlı olarak ilaç duyarlılıki tahmini çalışılmıştır. Çalışmanın ikinci aşamasında kanser hücre hatlarının ilaçlara karşı duyarlılığını tahmin etmeye yönelik makine öğrenmesi modelleri geliştirmek amacıyla çeşitli biyomedikal veri tabanlarından elde edilen açık erişimli veriler bütünleştirilmesini temel almıştır. GDCS (Genomics of Drug Sensitivity in Cancer), COSMIC (Catalogue Of Somatic Mutations In Cancer) ve DGIdb (Drug-Gene Interaction database) veritabanlarından farklı verilerin birleştirilmesi yoluyla ilaç duyarlılık analizi yapılmıştır.

3.1.Çalışmanın Birinci Aşaması ile İlgili Gereç ve Yöntem

Çalışmanın ilk aşamasında, ilaç duyarlılık ile mutasyon yükü arasındaki olası ilişkileri inceleyebilmek amacıyla farklı açık veri tabanlarından çeşitli biyolojik ve farmakolojik veriler derlenmiştir. Mevcut literatür ve kaynaklar incelendiğinde, mutasyon yükünü analiz için gerekli formatta doğrudan sunan bütüncül bir veri kümesine ulaşılamamıştır. Bu nedenle, farklı veri kaynaklarından elde edilen genomik bilgiler birleştirilerek mutasyon yükünün hesaplanabilir ve analiz edilebilir bir forma dönüştürülmesi sağlanmıştır. Bu süreçte hem veri uyumluluğu hem de biyolojik anlamlılık gözetilerek ön işleme adımları uygulanmıştır.

3.1.1. Veri toplama

Çalışmada, hücre hatları ve mutasyon profillerinin ilaç duyarlılığı üzerindeki etkisini incelemek amacıyla açık erişimli kaynaklardan faydalanılmıştır. Bu kapsamda öncelikle Genomics of Drug Sensitivity in Cancer (GDSC) veri tabanına ait Yang ve arkadaşları (2013) tarafından yayımlanan veri setleri incelenmiş ve indirilebilir formatta olan kaynaklar toplanmıştır. Hücre hatlarına ait mutasyon bilgileri, Cell Model Passports platformu üzerinden sağlanan kapsamlı "Mutasyonlar" veri setinden elde edilmiştir. Bu veri seti; her bir hücre hattı için gen

kimliği (gene_id), hücre modeli kimliği (cell_model_id), mutasyonun sürücü (driver) ya da kodlayıcı (coding) olup olmadığı, biyolojik etkisi ve varyant alel frekansı (VAF) gibi detaylı bilgileri içermektedir. Ancak bu kaynakta, hücre hatlarının ait olduğu kanser tiplerine dair bilgi bulunmamaktadır.

İlaçların, belirli mutasyon profiline sahip hücre hatları üzerindeki etkisinin analiz edilmesi çalışmanın odak noktalarından biri olduğundan, bu ilişkiyi kurabilmek adına ek veri kaynaklarına ihtiyaç duyulmuştur. Bu doğrultuda, GDSC veri tabanında yer alan yarı maksimal inhibitör konsantrasyonu (IC50) verisi incelenmiştir. Bu veri seti, farklı ilaçların çeşitli hücre hatlarında oluşturduğu yanıtları kantitatif olarak sunmakta olup, hücre-ilaç etkileşimlerinin analizinde temel bir referans noktası oluşturmuştur. Verilerin toplandığı tarihte, 265 farklı ilaç ya da ilaç adayı için IC50 değerleri mevcuttu.

Ayrıca, GDSC veri tabanındaki "hücre listesi" veri seti, her bir hücre hattının biyolojik ve histopatolojik özellikleriyle birlikte ait olduğu kanser alt tiplerini de içermektedir. Bu veri seti, hem hücre tanımlayıcıları hem de diğer veri kaynaklarıyla (mutasyon ve IC50 verileri) ortak anahtar değişkenler (örneğin hücre hattı ID'si) üzerinden entegre edilebilmesi açısından özel bir öneme sahiptir.

Çalışma kapsamında üç ana veri seti —mutasyon verisi, IC50 yanıt verisi ve hücre listesi— hücre listesi veri seti merkez alınarak birleştirilmiştir. Ancak bu entegrasyon sonucunda ortaya çıkan veri setinin oldukça yüksek sayıda gözlem içermesi, analizlerin anlamlı ve odaklı şekilde sürdürülebilmesi için filtreleme ve ön işleme adımlarının uygulanmasını zorunlu kılmıştır.

3.1.2. Lung Small Cell Carcinoma verisi

GDSC (Genomics of Drug Sensitivity in Cancer) veri tabanı, farklı kanser türlerine ait hücre hatlarında gözlemlenen ilaç duyarlılıki verilerini kapsamlı şekilde sunmaktadır. Bu çalışmada, analizlerin daha odaklı ve anlamlı biçimde yürütülebilmesi amacıyla sadece küçük hücreli akciğer kanseri (lung_small_cell_carcinoma) alt tipine ait hücre hatları ve bunlara ilişkin ilaç yanıt verileri değerlendirme kapsamına alınmıştır.

Veri setinin analiz öncesi işlenebilir hâle getirilebilmesi için, hücre listesi veri

seti içerisinde yer alan “subtype” özelliği filtrelenmiş ve yalnızca “lung_small_cell_carcinoma” olarak etiketlenen kayıtlar seçilmiştir. Bu filtreleme işlemi sayesinde hem biyolojik bağlamda homojen bir örneklem elde edilmiş hem de veri boyutu azaltılarak hesaplamaya uygun bir ölçeğe indirilmiştir. Bu ön işleme adımı sonucunda oluşturulan mutasyon veri seti, toplamda 512.464 gözlem (satır) içerecek şekilde yapılandırılmıştır.

3.1.3. Makine öğrenmesi için veri seti oluşturma

Analiz sürecinin sonraki aşamasında, makine öğrenmesi algoritmalarının uygulanabilmesi amacıyla yapısal ve sayısal biçimde uygun özellikler (feature engineering) tanımlanarak veri seti yeniden düzenlenmiştir. Bu amaç doğrultusunda, yalnızca küçük hücreli akciğer kanseri (Lung Small Cell Carcinoma) alt tipine ait hücre hatlarında gözlemlenen mutasyonlara ilişkin çeşitli biyolojik parametreler sayısallaştırılmış ve her bir hücre hattı–ilaç çifti için öznitelik temelli temsil yapısı oluşturulmuştur.

GDSC (Genomics of Drug Sensitivity in Cancer) veri tabanından türetilen bu yeni yapay veri setinde, her bir ilaç için bağımsız olarak aşağıda listelenen değişkenler hesaplanarak modele giriş değişkeni olarak tanımlanmıştır:

1. İlaç ismi,
2. Hücre isimleri,
3. Toplam mutasyon,
4. Toplam driver mutasyon,
5. Predisposition variant sayısı,
6. Mutasyon,
7. Toplam intronic mutasyon,
8. Toplam Frame Shift sayısı,
9. Start loss mutasyon sayısı,
10. Stop loss mutasyon sayısı,
11. Upstream mutasyon sayısı,
12. Nonsense mutasyon sayısı,
13. Kodlayıcı mutasyon sayısı,
14. VAF skoru,
15. İlaç duyarlılığı

3.1.4. Makine öğrenmesi uygulaması

Çalışma kapsamında, Sanger Enstitüsü tarafından sağlanan GDSC veri kaynakları kullanılarak, toplam beş farklı antikanser ilaca ilişkin mutasyon ve ilaç duyarlılığı verileri toplanmıştır. Seçilen ilaçlar şunlardır: bortezomib, vinorelbine, docetaxel, elesclomol ve ABT-263. Bu ilaçların belirlenmesinde iki temel ölçüt esas alınmıştır:

- ✓ IC50 değerleri açısından geniş varyasyon aralığına sahip olmaları, bu sayede duyarlılık spektrumunun makine öğrenmesi analizleri için anlamlı hale getirilmesi;
- ✓ Analiz yapılacak hücre hatlarının yalnızca küçük hücreli akciğer karsinomu (lung_small_cell_carcinoma) alt tipine ait olması.

Oluşturulan veri setinde, her bir hücre hattı için yukarıda belirtilen biyolojik öznitelikler (örneğin toplam mutasyon sayısı, driver mutasyon oranı, VAF ortalaması vb.) ile birlikte ilaç-hücre eşleşmelerine ait IC50 skorları da yer almaktadır. Böylece, her satırda belirli bir hücre hattı ile bu beş ilaçtan biri arasındaki duyarlılık ilişkisi sayısallaştırılmış biçimde temsil edilmiştir.

Analizlerde sadece lung_small_cell_carcinoma tanısı almış hücre hatları dikkate alınmış; tüm ilaçlara ait veriler birleştirilerek ortak bir dosyada bütünleştirilmiştir. Elde edilen nihai veri seti, her biri bir hücre-ilaca ait ilişkisel gözlemi temsil eden toplam 241 satırdan oluşmaktadır. Her bir satırda, mutasyon temelli açıklayıcı değişkenlerin yanı sıra, ilaç duyarlılığı çıktısı (IC50) da yer almakta olup, veri seti toplamda 14 sütun içermektedir. Bu yapı sayesinde, farklı ilaçların küçük hücreli akciğer kanseri hücre hatlarında gösterdiği etkiler, mutasyon profilleri temelinde karşılaştırmalı olarak değerlendirilmeye uygun hâle getirilmiştir.

3.1.5. Makine öğrenmesi modelleme

Oluşturulan ve toplamda 241 gözlem ile 14 öznitelikten oluşan veri seti, ilaç duyarlılığının öngörülebilmesi amacıyla makine öğrenmesi yöntemleriyle analiz edilmiştir. Bu veri setinde, ilaç duyarlılığı bağımlı değişken olarak tanımlanmış ve geri kalan özellikler açıklayıcı (bağımsız) değişkenler olarak kullanılmıştır. Modelleme sürecinin temel hedefi, her bir hücre–ilaç eşleşmesi için, hücrenin

mutasyon profiline dayalı olarak ilgili ilaca karşı gösterdiği duyarlılık düzeyini tahmin etmektir

Bu çalışmanın özgün yönü, farklı mutasyon yüklerinin toplam düzeyde ilaç duyarlılığı üzerindeki etkisinin sistematik olarak araştırılmasıdır. Elde edilen bulgular doğrultusunda, mutasyon yüküne ve bu yüke ait nitel özelliklere göre belirli ilaçlara karşı oluşabilecek direnç düzeyinin önceden tahmin edilebilmesi amaçlanmaktadır.

Modelleme sürecinde, küçük hücreli akciğer kanseri (SCLC) hücre hatlarına ait 61 farklı hücrenin yaklaşık 500.000 mutasyon kaydı dikkate alınarak analiz gerçekleştirilmiştir. Bu geniş veri kümesinden elde edilen mutasyon bilgileri, belirli bir oranla eğitim ve test alt kümelerine ayrılmış; modelin genelleme performansını değerlendirmek amacıyla eğitim kümesinden bağımsız hücre hatlarıyla test işlemi gerçekleştirilmiştir.

Makine öğrenmesi uygulamaları kapsamında; XGBoost, CatBoost ve Destek Vektör Makineleri (SVM) algoritmaları denenmiş ve farklı modelleme stratejileriyle bu yöntemlerin performansları karşılaştırılmıştır. Modellerin geliştirilmesi sürecinde, veri setinin %80'i eğitim, kalan %20'si test amacıyla kullanılmıştır. Bu ayrım yapılırken, hücre hattı isimleri temel alınmış; böylece her hücrenin bağımsız olarak test sürecine dahil edilebilmesi sağlanmıştır. Model doğrulama sürecinde kullanılan değerlendirme metriklerine göre, bazı ilaçlar için %60 düzeyinde R^2 (determinasyon katsayısı) elde edilmiştir; bu da modellenen değişkenin önemli bir kısmının açıklanabildiğini göstermektedir.

3.1.6. İdeal Model Parametre belirlenmesi

CatBoost algoritması için kullanılan parametreler aşağıdaki tabloda belirtilmiştir.

Tablo 3. CatBoost algoritması için kullanılan parametreler

Parametre	Değer
Seed	13
İterasyon	300
Learning rate	0,1

Derinlik	6
Loss function	RMSE
eval metric	RMSE

Bir diğer makine öğrenmesi modeli olan XG Boost algoritması ise benzer şekilde aşağıdaki başlangıç parametreleri ile çalıştırılmıştır.

Tablo 4. XG Boost algoritması için kullanılan parametreler

Parametre	Değer
max depth	6
nrounds	300
verbose	0
Learning rate	0,1

Çalışmanın son aşamasında, Destek Vektör Makineleri (Support Vector Machines, SVM) algoritması kullanılarak modelleme çalışmaları gerçekleştirilmiştir. SVM modeli için hiperparametre optimizasyonu amacıyla, farklı C (ceza parametresi) ve sigma (RBF çekirdek fonksiyonu genişliği) değerleri test edilmiştir. Bu kapsamda modelin genelleme performansını artırmak ve aşırı öğrenme (overfitting) riskini en aza indirmek için 10 katlı çapraz doğrulama (10-fold cross-validation) yöntemi uygulanmıştır.

Gerçekleştirilen parametre taramasında aşağıdaki değer aralıkları kullanılmıştır:

- ✓ C: 0,25, 0,5, 1
- ✓ sigma: 0,1, 0,01, 0,001

Her bir parametre kombinasyonu için elde edilen performans ölçütleri (örneğin R^2, MAE, RMSE gibi) karşılaştırılarak en iyi sonuç veren yapılandırmalar belirlenmiştir. Bu süreçte kullanılan algoritmik yaklaşım, model doğruluğunu değerlendirmenin yanı sıra, parametrelerin model çıktısı üzerindeki etkilerini de analiz etmeyi mümkün kılmıştır.

Tüm uygulamalar R programlama dili kullanılarak gerçekleştirilmiş, model eğitim ve test işlemleri Intel i7 7. nesil işlemciye ve 16 GB RAM'e sahip bir bilgisayarda yürütülmüştür. Bu donanım özellikleri, orta ölçekli biyomedikal veri kümeleri için yeterli işlem kapasitesi sağlamıştır.

Deneysel çalışmalarda kullanılan yöntemler, yeniden üretilebilirlik ilkesine

uygun olarak detaylı biçimde belgelenmiş ve her adım açıklanabilir hale getirilmiştir. Bu bağlamda, çalışmanın hem doğrulanabilir hem de benzer araştırmalarda uygulanabilir nitelikte olması hedeflenmiştir. Teorik ya da analitik bileşenlerin yer aldığı durumlarda ise, kullanılan formüller literatür kaynaklarına dayandırılarak sunulmuş; gerekli durumlarda kısaca türetilmiş ve formüllerde geçen tüm büyüklükler tanımlanmıştır.

3.2. Çalışmanın İkinci Aşaması ile İlgili Gereç ve Yöntem

3.2.1. Veri setlerinin birleştirilmesi ve ön işleme süreci

Çalışmada ilk olarak, GDSC, COSMIC ve DGIdb veri tabanlarından elde edilen açık kaynaklı veriler birleştirilmiştir. Daha sonra, bu entegre veri seti üzerinde filtreleme, metin tabanlı kodlama ve meta-sezgisel yöntemlerle özellik seçimi gerçekleştirilmiştir. Bu işlemler sonucunda elde edilen veri kümesi, makine öğrenmesi algoritmalarına girdi olarak kullanılmıştır. Son aşamada ise makine öğrenmesi modelleri entegre veri ile eğitilmiş, parametreleri optimize edilmiş ve geliştirilen model ilaç duyarlılığını öngörmek amacıyla Z-skorunu tahmin edecek şekilde tasarlanmıştır.

3.2.2. Açık erişimli veri kaynaklarının toplanması ve entegrasyonu ve GDSC verilerinin toplanması

Çalışmanın başlangıcında GDSC, COSMIC ve DGIdb veri tabanlarından elde edilen dört farklı veri seti bir araya getirilmiştir. Öncelikle ilaç duyarlılığına ilişkin veriler GDSC'den alınmış, ardından COSMIC veri tabanından elde edilen iki ayrı dosya GDSC ile entegre edilerek somatik mutasyonlar ile ilaç yanıtlarının eşzamanlı değerlendirilmesi sağlanmıştır. Son aşamada, gen–ilaç etkileşimlerini içeren DGIdb verileri dahil edilerek analiz zenginleştirilmiştir. İlaç duyarlılığını tahmin etmek amacıyla GDSC1 ve GDSC2 veri setleri kullanılmıştır. Bu veri setleri, farklı kanser hücre hatlarının çeşitli ilaçlara verdiği yanıtları yansıtan genetik parametreleri kapsamaktadır. Birleştirilmiş dosya;

NLME_RESULT_ID,
COSMIC_ID,
CELL_LINE_NAME,
DRUG_NAME,
PUTATIVE_TARGET,
PATHWAY_NAME,
LN_IC50,
AUC,
RMSE ve
Z_SCORE gibi değişkenleri içermektedir. Veri setlerinin entegre edilmesi, hem veri miktarını artırmış hem de farklı deney koşullarından elde edilen bilgilerin çeşitlenmesini sağlamıştır. Böylece, daha güçlü ve genellenebilir makine öğrenmesi modelleri oluşturulabilmiştir.

3.2.3. GDSC ve COSMIC verilerinin entegrasyonu

İkinci adımda, COSMIC veri tabanındaki CellLinesProject_CompleteCNA_v101_GRCh38.tsv.gz dosyası GDSC verileri ile birleştirilmiştir. Bu veri seti aşağıdaki sütunları içermektedir:
COSMIC_CNV_ID,
COSMIC_GENE_ID,
COSMIC_SAMPLE_ID,
SAMPLE_NAME,
COSMIC_PHENOTYPE_ID,
TOTAL_CN,
MINOR_ALLELE,
MUT_TYPE,
COSMIC_STUDY_ID,
CHROMOSOME,
GENOME_START,
GENOME_STOP ve
GENE_SYMBOL sütunları. COSMIC veri setindeki COSMIC_SAMPLE_ID

sütunu, GDSC veri setindeki COSMIC_ID sütunu ile eşleştirilerek entegrasyon yapılmıştır.

COSMIC Hücre Hattı Projesi'nin veri setine eklenmesi, veri çeşitliliğini artırmıştır. Nitekim, her hücre hattına ilişkin kopya sayısı değişikliği bilgisi farmakolojik verilerle eşleştirilmiş, bu sayede genetik değişimlerin ilaç etkinliği üzerindeki olası etkilerinin incelenmesine olanak tanınmıştır.

3.2.4. DGIdb'nin entegrasyonu

Üçüncü adımda, DGIdb veri tabanından elde edilen interactions.tsv dosyası kullanılarak veri seti, gen–ilaç etkileşim bilgileriyle zenginleştirilmiştir. Bu dosyadan
GENE_CLAIM_NAME,
GENE_CONCEPT_ID,
GENE_NAME,
INTERACTION_SOURCE_DB_NAME,
INTERACTION_SOURCE_DB_VERSION,
INTERACTION_TYPE,
INTERACTION_SCORE,
DRUG_CLAIM_NAME,
DRUG_CONCEPT_ID,
DRUG_NAME,
APPROVED,
IMMUNOTHERAPY ve
ANTI_NEOPLASTIC sütunları eklenmiştir. Eşleştirme süreci, DGIdb'deki GENE_NAME ve DRUG_NAME sütunları ile mevcut veri setindeki GENE_SYMBOL ve DRUG_NAME sütunlarının karşılaştırılması yoluyla gerçekleştirilmiştir. Bu yöntem, biyolojik açıdan anlamlı gen–ilaç etkileşimlerini ortaya çıkarmış, özellikle de onkolojiye yönelik potansiyel tedavi hedeflerini belirlemiştir. Ancak, veri tabanları arasındaki adlandırma farklılıkları sebebiyle bazı kayıtların dışlanması gerekmiştir.

3.2.5. COSMIC genom tarama mutasyonlarının entegrasyonu

Çalışmanın son aşamasında, COSMIC veri tabanında yer alan CellLinesProject_GenomeScreensMutant_v101_GRCh38.tsv.gz dosyası mevcut veri setine eklenmiştir. Bu dosyadaki somatik mutasyon bilgileri, COSMIC_SAMPLE_ID ve GENE_SYMBOL sütunları kullanılarak entegre veri ile eşleştirilmiştir. Çalışmada ileri değerlendirme için şu sütunlar kullanılmıştır:
COSMIC_SAMPLE_ID,
GENE_SYMBOL,
COSMIC_GENE_ID,
TRANSCRIPT_ACCESSION,
MUTATION_ID,
MUTATION_DESCRIPTION,
MUTATION_AA,
MUTATION_CDS,
LOH,
HGVSP,
HGVSC,
GENOMIC_MUT_ALLELE,
GENOMIC_WT_ALLELE,
MUTATION_SOMATIC_STATUS.

Böylece, hücre hatlarında saptanan belirli mutasyonların ilaç duyarlılığı üzerindeki etkilerinin daha kapsamlı bir biçimde incelenmesi mümkün olmuştur. Bu işlemler sonucunda, hem farmakolojik hem de genetik bilgileri barındıran yüksek boyutlu bir bütünleşik veri seti oluşturulmuştur. Söz konusu veri seti, potansiyel ilaç hedeflerinin belirlenmesi ve Z_SCORE gibi ilaç duyarlılığı ölçütlerinin öngörülmesi için sağlam bir altyapı sunmaktadır.

3.2.6. Veri ön işleme ve özellik seçimi

Oluşturulan entegre veri setinin doğrudan makine öğrenmesi modeli geliştirmeye uygun olmadığı tespit edilmiştir. Bunun nedeni, verilerin karmaşık yapısı ve olası

tekrar kayıtlarıdır. Bu nedenle, gereksiz ve yinelenen değişkenler temizlenmiş, yalnızca bilgi değeri yüksek olan değişkenler korunmuştur. Ayrıca, eksik verilerin tamamlanması da gerekli olmuştur. Literatürde, dört ayrı veri setinin bu şekilde bütünleştirilmesine yönelik özel bir aracın daha önce geliştirilmediği bilinmektedir.

3.2.7. Kategorik verilerin kodlanması ve eksik alanların düzenlenmesi

Entegre veri setinde yer alan kategorik değerler, kardinalitelerine uygun biçimde sayısallaştırılmıştır. Sayısal modellemeye katkısı olmayan yüksek kardinaliteli sütunlar (örneğin genomik konum ve spesifik allel bilgileri) dışlanmıştır. Kodlama işlemlerinde farklı stratejiler uygulanmıştır: düşük kardinaliteli değişkenler için One-Hot Encoding, orta kardinaliteli değişkenler için Target Encoding, yüksek kardinaliteli değişkenler için ise Ordinal Encoding tercih edilmiştir.

3.2.8. Eksik verilerin tamamlanması ve yanlılığın azaltılması

Sayısal verilerdeki eksik alanlar, istatistiksel yöntemlerle tamamlanarak dengeli bir veri yapısı oluşturulmuştur. Böylelikle modelin eğitim ve doğrulama süreçlerinde hem doğruluk artmış hem de yanlılık azaltılmıştır. Makine öğrenmesine uygun bir dosya elde etmek için, öncelikle kapsamlı bir ön işleme ve değişken seçimi süreci yürütülmüştür.

Bu süreçte, veri birleşiminden kaynaklanan tekrar sütunlar (ör. GENE_NAME ile GENE_CLAIM_NAME veya _x/_y uzantılı sütunlar) kaldırılmış; modelin öngörü gücüne katkı sağlamayan INSTITUTE, WEBRELEASE, COMPANY_ID gibi meta veriler de çıkarılmıştır.

Çalışmada hedef değişken olarak Z_SCORE belirlenmiş, bu değişkenle yüksek korelasyon gösteren RMSE, AUC ve LN_IC50 gibi ölçütler modelde yanlı tahminlere yol açmamak için ayrıştırılmıştır. Son olarak, hedef değişken ile açıklayıcı değişkenler birbirinden ayrılmış ve makine öğrenmesi modellerinin eğitim ve test süreçlerine uygun temiz bir veri seti oluşturulmuştur.

3.2.9. Drug–Gene etkileşimlerinin keşifsel incelemesi

Çalışmada ilaç etkinliğini ölçmek amacıyla Z-skorları kullanılmıştır. Negatif Z-skoru, ilgili ilacın hedef gene güçlü bir baskılayıcı etki yaptığını; pozitif skor ise etkinliğin düşük olduğunu göstermektedir (Yang ve ark., 2013). Entegre veri seti özelinde Z-skorları, özellikle kanser tedavisinde hedeflerin önceliklendirilmesinde yol gösterici bir parametre olarak değerlendirilmiştir.

Veri entegrasyonu tamamlandıktan sonra, ilaç–gen ilişkilerini ortaya koymak üzere keşifsel analizler yapılmıştır. Analizlerde, her bir ilacın hedeflediği gen sayısı ile her genin ilişkilendirildiği ilaç sayısı hesaplanmıştır. Ayrıca, etkileşim yoğunluğu incelenmiş ve bu yoğunluk toplam etkileşim sayısı üzerinden değerlendirilmiştir.

3.2.10. İlaç ve genlerin kümeleme analizleri

Gen ekspresyonu ve ilaç yanıtlarına göre ilaçları gruplandırmak için Ward'ın minimum varyans yöntemi uygulanmıştır (Ward, 1963). İlaçlar ve genler arasındaki benzerlikler, aktivite profilleri temel alınarak hiyerarşik kümeleme yöntemleri ile değerlendirilmiştir. Bu süreçte, kümeler arasındaki toplam iç varyansı azaltmaya yönelik ardışık birleştirme yaklaşımı benimsenmiş ve uzaklık ölçütü olarak Öklidyen mesafe kullanılmıştır.

Aynı yöntem genler için de uygulanmış; genlerin ilaç yanıt profilleri esas alınarak kümeler oluşturulmuştur. Böylece, benzer yanıt özelliklerine sahip gen grupları belirlenebilmiştir. İlaçlar ve genler için elde edilen kümeler, ısı haritalarıyla görselleştirilmiş ve Z-skorlarıyla normalize edilmiştir. Ek olarak, ilaç–gen ilişkilerini daha iyi anlamak amacıyla Pearson korelasyon katsayılarına dayalı korelasyon matrisleri hesaplanmış ve bu sonuçlar da ısı haritası biçiminde sunulmuştur (Benesty, Chen, Huang, & Cohen, 2009). Tüm bu analizler, yüksek boyutlu genetik verilerdeki örüntüleri ortaya koymak amacıyla sistematik olarak yürütülmüştür.

Sonuçta, yürütülen işlem hattının tüm adımlarıyla oluşturulan manipüle edilmiş veri seti, dört farklı kaynaktan elde edilen özellikleri barındırmakta olup, 26 bağımsız ve 1 bağımlı değişken içermektedir.

3.2.11. Modelleme öncesi özellik seçiminde optimizasyon

Bölüm 2.7'de belirtildiği üzere, veri işleme sonrası elde edilen veri seti 27 özelliğe indirgenmiştir. Özellik seçiminin nihai aşamasında, yaklaşık 27 aday değişken arasından en bilgilendirici alt kümeyi belirleyerek modelin tahmin gücünü en üst seviyeye çıkarmak için Parçacık Sürüsü Optimizasyonu (PSO) algoritması kullanılmıştır (Kennedy, & Eberhart, 1995). PSO, geniş arama alanlarını etkili biçimde tarayabilmesi, küresel optimuma yakın çözümler bulabilmesi ve yüksek boyutlu optimizasyon problemlerinde yerel tuzaklara karşı dayanıklı olması nedeniyle bu çalışma için güçlü bir yöntem olarak tercih edilmiştir.

Bu süreçte 40 parçacık ve 500 iterasyon içeren deneyler gerçekleştirilmiştir. Seçilen özellik alt kümesinin başarısı R^2 skoru ile değerlendirilmiş, ayrıca Random Forest regresyon modeli (Breiman, 2001) kullanılarak beş katlı çapraz doğrulama yapılmıştır. R^2 değeri, PSO temelli özellik seçiminin uygunluk kriteri olarak kullanılmıştır. Elde edilen sonuçlar, yinelenen veya gereksiz özelliklerin elenmesine ve yalnızca bilgi değeri yüksek olanların korunmasına imkân vermiştir. Sonuçta, en anlamlı 11 özellik belirlenmiştir.

PSO ile gerçekleştirilen optimizasyon, CELL_LINE_NAME, SANGER_MODEL_ID, TCGA_DESC, DRUG_ID, PUTATIVE_TARGET, PATHWAY_NAME, TOTAL_CN, MINOR_ALLELE, interaction_score, interaction_type_missing ve LOH_u özelliklerinin Z-skoru tahmininde en uygun değişkenler olduğunu göstermiştir. Böylelikle özellik uzayı daraltılmış, ancak modelin doğruluk ve genelleme kabiliyetinde herhangi bir kayıp yaşanmamıştır.

3.2.12. Son veri seti ve performans ölçütleri

Çalışmada kullanılan veri seti başlangıçta 5.747 örnek ve 72 özellikten oluşmaktadır. Filtreleme işlemleri tamamlandıktan sonra özellik sayısı 27'ye düşürülmüştür. Bu değişkenlerden 26'sı giriş verisi olarak kullanılırken, Z-skoru çıkış (yanıt) değişkeni olarak belirlenmiştir. Modelin Z-skoruna ilişkin tahmin gücünün artışı, R^2 değerindeki yükselme ile ortaya konmuştur.

Elde edilen 27 özellik; hedef genler, biyolojik yolaklar, mutasyon bilgileri ve

diğer parametreleri kapsayan karma bir yapıdadır. Algoritmanın nihai aşamasında ise, bu 27 özellik arasından en bilgilendirici olan 11 özellik seçilmiştir.

3.2.13. Makine öğrenmesi modellerinin oluşturulması

Bu aşamada çalışmanın temel hedefi, makine öğrenmesi algoritmaları aracılığıyla ilaç duyarlılığını en doğru şekilde öngörmektir. Farklı algoritmalar kullanılarak yeni modeller geliştirilmiş ve bu modeller doğruluk düzeyleri bakımından karşılaştırılmıştır.

3.2.14. Veri setinin bölünmesi

Model performansını ölçmek için veri seti eğitim ve test kümelerine ayrılmıştır. train_test_split yöntemi kullanılarak verinin %80'i eğitim, %20'si test için ayrılmıştır. Böylelikle, model eğitim verisi üzerinde öğrenmiş ve bağımsız test verisiyle doğrulanarak genel tahmin performansı ölçülmüştür.

3.2.15. Özellik seçimi ve algoritmaların uygulanması

Model doğruluğunu yükseltmek ve aşırı öğrenmeyi önlemek amacıyla veri seti üzerinde özellik seçimi yapılmıştır. İlk adımda, eğitim veri setinde değişkenliği bulunmayan (sabit değerli) değişkenler VarianceThreshold yöntemiyle veri setinden çıkarılmıştır. Böylelikle, tüm örneklerde aynı değeri alan ve model performansına katkı sağlamayan değişkenler elenmiştir.

Devamında, bağımsız değişkenler ile bağımlı değişken (Z_SCORE) arasındaki ilişkiyi belirlemek için f_regression testi kullanılmıştır. Bu analizde her bir değişkenin varyans katkısı ve anlamlılık düzeyi (p-değeri) hesaplanmış; $p<0,05$ değerine sahip değişkenler modele dâhil edilmiştir. Eğer anlamlı bir değişken bulunmazsa, model mevcut tüm değişkenlerle eğitilmeye devam etmiştir. Model geliştirme sürecinde kullanılan algoritmalar şunlardır: Random Forest, Extra Trees, LightGBM, CatBoost ve XGBoost.

Seçilen 11 özellik ile beş farklı regresyon modeli geliştirilmiştir: XGBoost

(Chen, & Guestrin, 2016), CatBoost (Prokhorenkova ve ark., 2018), LightGBM (Ke ve ark., 2017), Random Forest (Breiman, 2001) ve Extra Trees (Geurts, Ernst, & Wehenkel, 2006). Bu ağaç tabanlı yöntemler, eksik veriyle çalışabilme, yüksek boyutlu biyolojik verileri yorumlayabilme ve güçlü genelleme performansları nedeniyle tercih edilmiştir.

Her algoritmada en uygun parametrelerin belirlenmesi için hiperparametre optimizasyonu yapılmış, bu amaçla RandomizedSearchCV yöntemi uygulanmıştır (Bergstra, & Bengio, 2012). Sistematik olarak test edilen parametreler şunlardır: XGBoost: n_estimators, max_depth, learning_rate, colsample_bytree, CatBoost: depth, learning_rate, l2_leaf_reg, LightGBM: n_estimators, num_leaves, learning_rate, Random Forest ve Extra Trees: n_estimators, max_depth, max_features.

Bu optimizasyon süreci sonucunda, gereksiz özelliklerden arındırılmış, daha sade yapıda ve yüksek doğruluk sağlayan modeller geliştirilmiştir.

4. BULGULAR

Tez çalışmamın bulgular kısmı, çalışmanın ilk aşaması ve çalışmanın ikinci aşaması olarak iki kısımdan oluşmaktadır.

4.1. Çalışmanın Birinci Aşaması ile İlgili Bulgular

Deneysel çalışma iki temel aşamadan oluşmaktadır. İlk aşamada, küçük hücreli akciğer kanseri (SCLC) hücre hatlarına ait mutasyon yüküne dayalı öznitelikler oluşturulmuş ve modellemeye uygun bir veri seti yapılandırılmıştır. Bu aşamada, mutasyonların türü, sayısı, etkisi ve varyant alel frekansı gibi biyolojik özellikler sayısallaştırılarak her hücre hattı için tanımlayıcı nitelikler elde edilmiştir.
İkinci aşamada ise, oluşturulan bu öznitelikler kullanılarak ilaç duyarlılığının (IC50 değeri) tahmini amaçlanmıştır. Bu doğrultuda, çeşitli makine öğrenmesi algoritmaları geliştirilmiş ve bu modellerin tahmin başarımı farklı istatistiksel performans ölçütleri üzerinden değerlendirilmiştir.

Model performansını nicel olarak değerlendirebilmek amacıyla üç temel metrik kullanılmıştır:

- ✓ RMSE (Root Mean Squared Error): Ortalama hata karelerinin karekökü
- ✓ R^2 (Determination Coefficient): Modelin açıklayıcılık oranı
- ✓ MAE (Mean Absolute Error): Ortalama mutlak hata değeri

Her bir makine öğrenmesi modelinin bu ölçütler üzerinden performansı karşılaştırılmış ve elde edilen sonuçlara göre en uygun model belirlenmiştir. Bu sayede, mutasyon yükü temelli özniteliklerin ilaç duyarlılığı üzerindeki belirleyiciliği sistematik olarak analiz edilmiştir.

4.1.1. Mutasyon yük veri üretmenin zorluğu

Çalışmada kullanılan veri setlerinden biri, hücre hatlarında yer alan tüm mutasyonları içeren ve yaklaşık 893 MB boyutunda olan kapsamlı mutasyon veri dosyasıdır. Hücrelere ilişkin tanımlayıcı bilgileri barındıran Hücre Listesi dosyası ise

yaklaşık 128 KB büyüklüğündedir. Bu iki veri kaynağının doğrudan kartezyen çarpımı (her bir hücreye karşı tüm mutasyonların eşleştirilmesi) kullanılarak birleştirilmesi, mevcut bilgisayar belleğinin sınırlı olması nedeniyle ciddi performans sorunlarına ve bellek yetersizliğine yol açma potansiyeline sahiptir.

Bu nedenle, veri işleme sürecinde öncelikli olarak filtreleme işlemi gerçekleştirilmiştir. Özellikle, yalnızca küçük hücreli akciğer karsinomu (SCLC) alt tipine ait hücre hatları veri setinden seçilmiş ve böylece veri boyutu önemli ölçüde azaltılmıştır. Bu adım, sonraki aşamalarda daha verimli veri manipülasyonu yapılabilmesi açısından kritik öneme sahiptir.

İlaç-hücre etkileşimlerine ait duyarlılık ölçümlerini içeren IC50 veri dosyası, yaklaşık 3.160 KB boyutundadır. Filtrelenmiş hücre verileri ile bu IC50 verilerinin birleştirilmesi hedeflenmiş; sonuçta, hem seçilen ilaçlara ait ilaç duyarlılık skorları hem de bu skorlara karşılık gelen hücre biyolojik özellikleri tek bir yapı altında bütünleştirilmiştir.

Ancak, 8 GB RAM kapasitesine sahip standart bir bilgisayar sisteminde, tüm ham verilerin doğrudan işlenmesi sırasında bellek yetersizliği yaşanmıştır. Bu nedenle, veri işleme sırasında bellek verimliliğini maksimize edebilmek amacıyla filtreleme işlemlerinin erken aşamalarda uygulanması zorunlu hale gelmiştir. Böylece, büyük boyutlu veri kümeleri ile çalışırken karşılaşılan sistem kaynaklı darboğazlar en aza indirgenmiş ve işlem süreleri optimize edilmiştir.

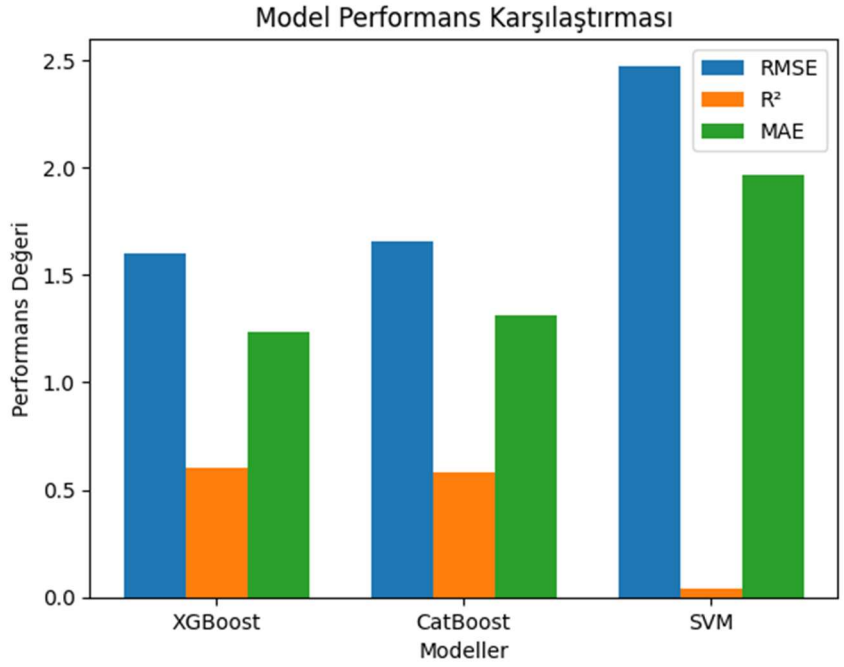
4.1.2. Makine öğrenme algoritma ile tahminler

Bu çalışmada, kanserli hücre hatları ile seçilmiş antikanser ilaçlar arasındaki IC50 değerlerinin tahmin edilmesi amacıyla farklı makine öğrenmesi algoritmaları kullanılmıştır. Modelleme sürecinde öznitelikler olarak hücrelerin mutasyon yüküne ilişkin sayısal veriler, hedef değişken olarak ise her hücre–ilaç eşleşmesi için IC50 değeri kullanılmıştır.

Uygulanan algoritmalar XGBoost, CatBoost ve Destek Vektör Makineleri (SVM) olup, her bir model için performans ölçütleri ayrı ayrı hesaplanmış ve karşılaştırılmıştır. Modellerin tahmin başarımı üç temel metrik üzerinden değerlendirilmiştir:

- ✓ R^2 (Determinasyon Katsayısı)
- ✓ RMSE (Kök Ortalama Kare Hatası)
- ✓ MAE (Ortalama Mutlak Hata)

Bu kapsamda geliştirilen makine öğrenmesi modellerine ilişkin performans sonuçları Şekil 1'de özetlenmiştir. Tablo her algoritmanın test verisi üzerindeki tahmin gücünü nicel olarak karşılaştırmalı biçimde sunarak, en uygun model seçimini desteklemeyi amaçlamaktadır.



Şekil 1. XGBoost, CatBoost ve SVM algoritmaları baz alınarak geliştirilen modeller

Şekil 1'de sunulan sonuçlar doğrultusunda, test verileri üzerinde en yüksek tahmin performansını sergileyen algoritmanın XGBoost olduğu görülmüştür. Bu model, R^2, RMSE ve MAE ölçütleri bakımından diğer algoritmalara kıyasla daha başarılı sonuçlar vermiştir. Bu durum, gradyan artırmalı karar ağaçları (gradient boosting) temelli algoritmaların, kanser hücrelerinde ilaç duyarlılığını tahmin etme konusunda güçlü bir alternatif sunduğunu göstermektedir.

Çalışma kapsamında test edilen örneklem üzerinde XGBoost modeli, ilaç–kanserli hücre etkileşimini deneysel testlere gerek kalmadan yaklaşık 1.60 RMSE, 0.60 R^2 ve 1.23 MAE oranlarında doğrulukla öngörebilmiştir. Bu seviyedeki tahmin başarımı, özellikle mutasyon profiline göre hangi ilaçların etki potansiyeli taşıdığına

dair ön bilgi sunması açısından klinik ve biyoinformatik bağlamda önem arz etmektedir.

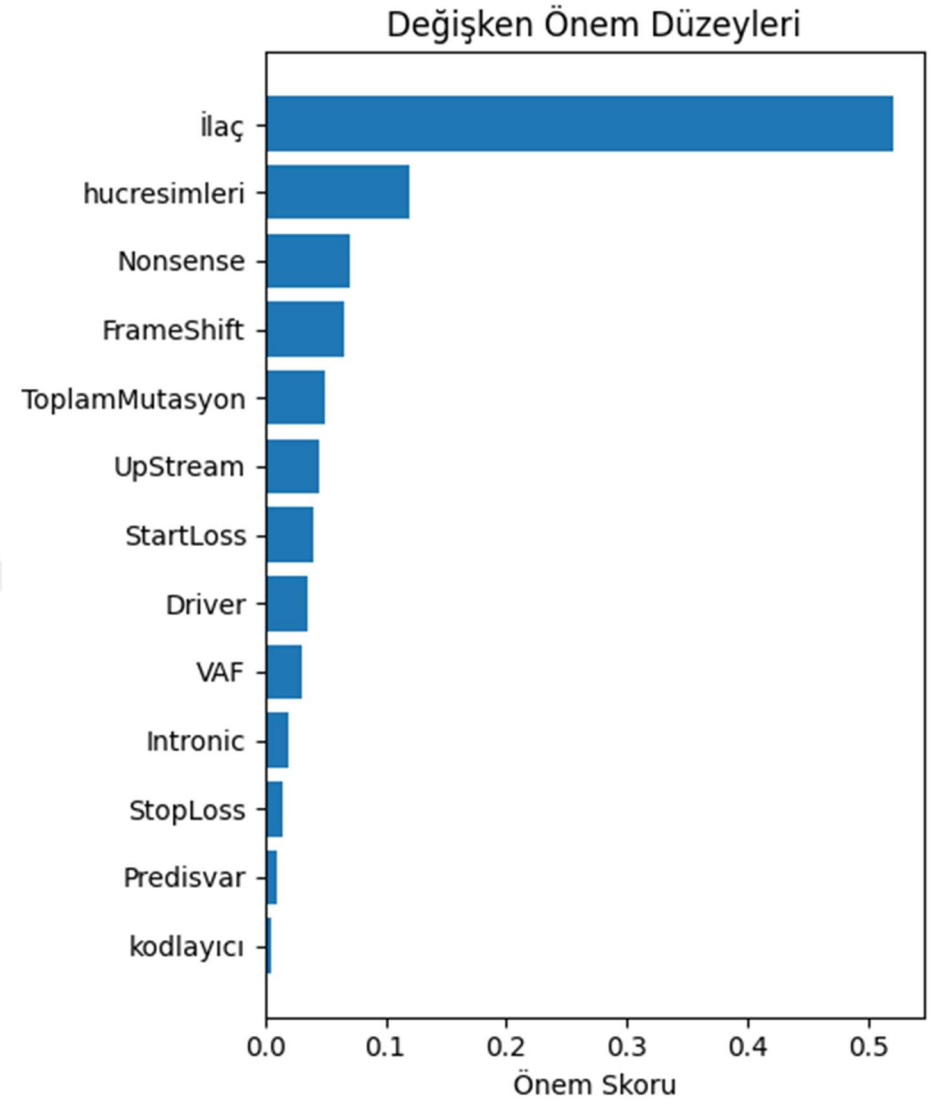
Modelin karar mekanizmasının daha iyi anlaşılabilmesi amacıyla, XGBoost tarafından öznitelik önem dereceleri aşağıda sıralanmıştır. Eğitim sürecinde verilerin %80'i modele sunulduğundan, modelin ilaç ve hücre adı gibi tanımlayıcı bilgileri öğrenmiş olması beklenmektedir. Ancak bu temel bilgilerin ötesinde, modelin kararında en etkili değişkenlerin sırasıyla:

- ✓ Nonsense mutasyon sayısı,
- ✓ Frameshift mutasyon sayısı,
- ✓ Toplam mutasyon yükü,
- ✓ Upstream mutasyon sayısı ve
- ✓ Start-loss mutasyon sayısı olduğu gözlemlenmiştir.

İlginç bir biçimde, bu özniteliklerin driver mutasyon sayısından daha fazla ağırlığa sahip olması, modelin kanser hücrelerinin genel genetik bozunum düzeyini (genomic instability) daha belirleyici bir faktör olarak değerlendirdiğini düşündürmektedir.

Bu bağlamda, yukarıda belirtilen mutasyon türlerinin sayısal olarak artması, belirli ilaçlara karşı hücresel duyarlılığın gelişme olasılığını artırabilir. Bu bulgu, ilaç seçimi öncesinde genetik profilin ayrıntılı şekilde değerlendirilmesinin önemine işaret etmektedir.

4.1.3. Mutasyon türlerinin etkisi ve model karmaşıklık analiz



Şekil 2. XGBoost Model grafiği

Elde edilen bulgular doğrultusunda, tahmin performansının en yüksek olduğu algoritma olarak öne çıkan XGBoost modeli ile ilgili teknik yapıya dair daha ayrıntılı bilgi sunulması uygun görülmüştür. XGBoost, gradyan artırmalı karar ağaçları mantığına dayanan güçlü bir topluluk öğrenme algoritması olup, yüksek boyutlu ve heterojen biyolojik verilerde öngörü doğruluğunu artırmak amacıyla sıklıkla tercih edilmektedir.

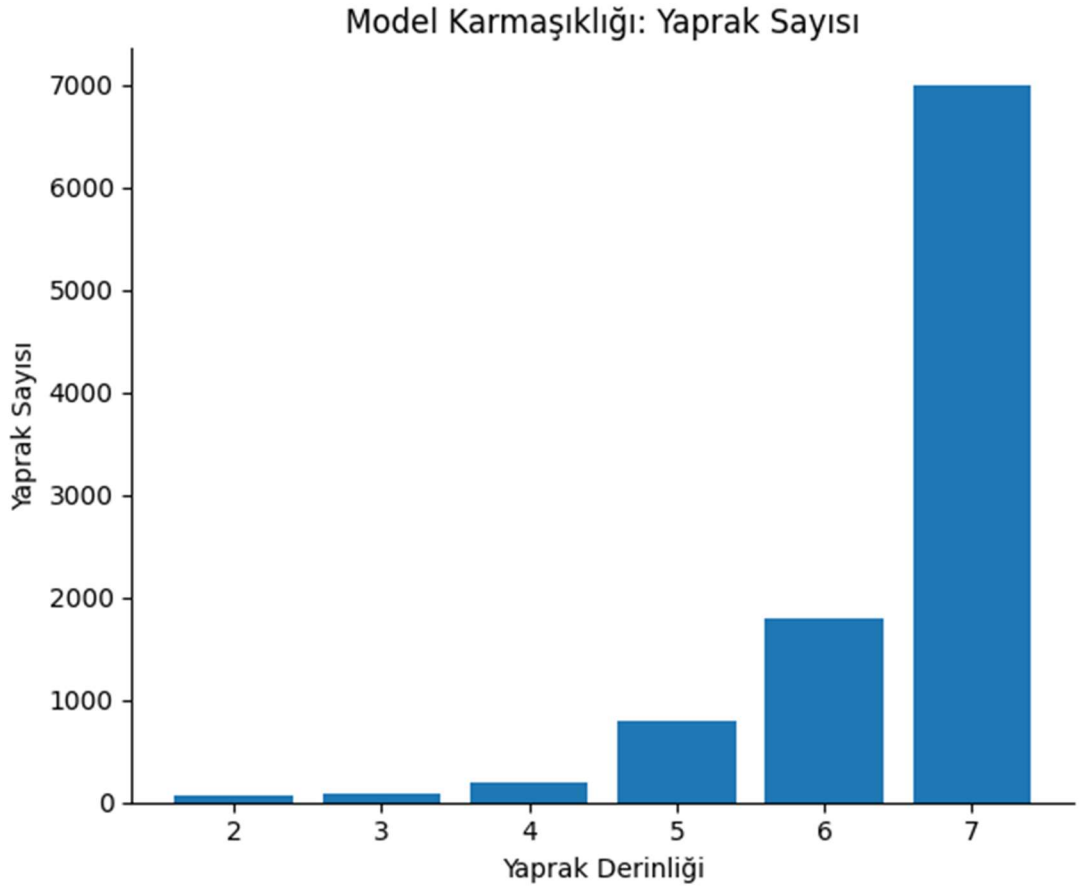
Modelin karar verme sürecinde hangi özniteliklerin hangi düzeyde katkı sağladığını anlamanın ötesinde, model karmaşıklığının da değerlendirilmesi önem arz etmektedir. Bu kapsamda, modelin karmaşıklık derecesini, yani karar ağaçlarının derinliği, sayısı ve genel yapı yoğunluğunu gösteren XGBoost Model Karmaşıklığı

grafiği, Şekil 2'de sunulmuştur.

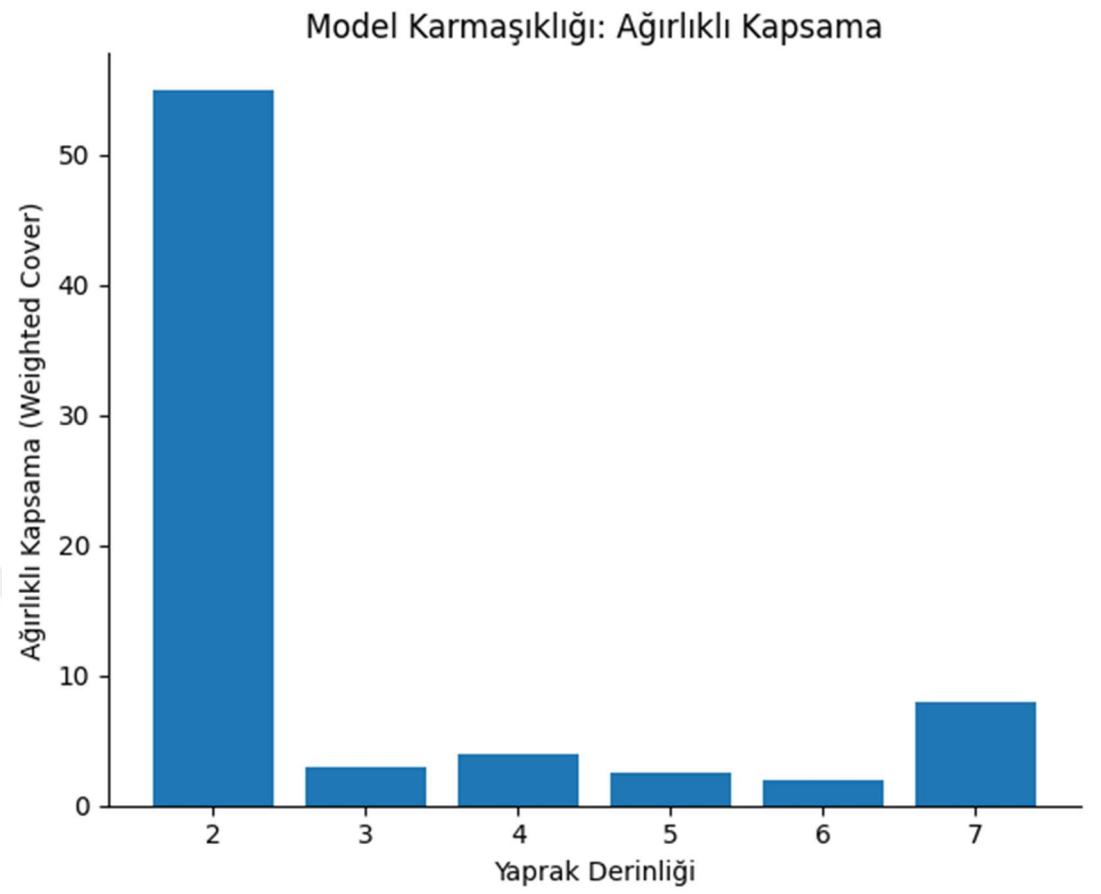
Bu grafik, modelin ne ölçüde kompleks yapılar inşa ettiğini ve öğrenme sürecinde aşırı uyum (overfitting) riskinin nasıl yönetildiğini ortaya koymak amacıyla oluşturulmuştur. XGBoost'un temel avantajlarından biri, düzenleme (regularization) parametreleri sayesinde modelin karmaşıklığını dengeleyebilmesidir. Bu sayede model hem yüksek öngörü performansı sunabilmekte hem de genelleme yeteneğini koruyabilmektedir.

Şekil 3)A üzerinden yorumlandığında, modelin eğitim sırasında sınırlı derinlikte ve kontrollü sayıda karar ağacı kullandığı görülmektedir. Bu yapı, veri setinin boyutu ve yapısıyla dengeli bir ilişki içinde olup, elde edilen tahmin başarısının altında yatan teknik altyapıyı yansıtmaktadır.

A)



B)



Şekil 3A/B. Yaprak derinliğine bağlı olarak XGBoost ağacında ağırlıklı kapsam ve yaprak sayısı grafikleri

Şekil 3)A ve 3)B' de, XGBoost algoritması tarafından oluşturulan karar ağaçlarında, yaprak derinliği ile ilişkili olarak ağırlıklı kapsam (weighted cover) ve yaprak sayısı arasındaki ilişki görselleştirilmiştir. Bu analiz, modelin sınıflandırma kararlarını hangi derinlik seviyelerinde yoğunlaştırdığını ve hangi derinliklerin daha belirleyici olduğunu incelemek amacıyla gerçekleştirilmiştir.

Weighted cover metriği, her bir karar ağacında öğrenme sürecinde ağırlıklı olarak kullanılan veri noktalarının toplamını temsil eder ve modelin hangi bölümlerinde daha fazla bilgi taşındığını gösterir. Bu bağlamda, ağacın belirli derinliklerindeki karar bölümlerinin model çıktısına katkı düzeyi hakkında önemli ipuçları sunar.

Şekil 3)A ve 3)B' deki dağılım incelendiğinde, özellikle 2. ve 7. seviye ağaç derinliklerinin sınıflandırma üzerinde etkili olduğu gözlemlenmektedir. Bu durum, modelin karar sürecinde hem erken ayrımlar (sığ derinlikte genel kurallar) hem de daha detaylı ayrımlar (derin ağaç yapıları) kullandığını göstermektedir. Özellikle 7.

derinlik seviyesine kadar çıkarılan karar ağaçlarının sınıflandırma performansına anlamlı katkı sağladığı izlenmektedir.

Bu bulgu, XGBoost'un derinlik parametresinin (max_depth) doğru ayarlanmasının, model başarımı üzerinde doğrudan etkili olduğunu ortaya koymaktadır. Derinliğin çok düşük tutulması modelin öğrenme kapasitesini sınırlandırırken, optimal derinlik düzeyine ulaşmak, hem genelleme gücünü artırmakta hem de sınıflandırma doğruluğunu optimize etmektedir.

4.2. Çalışmanın İkinci Aşaması ile İlgili Bulgular

4.2.1. Veri entegrasyonu ve deneysel tasarım

Çalışmada, ilaç duyarlılığı tahminlerinin doğruluğunu artırmak amacıyla dört farklı veri kaynağı bir araya getirilmiştir. Öncelikle GDSC1 ve GDSC2 veri setleri birleştirilmiş, ardından bu veri setleri COSMIC veri tabanından alınan CNA ve mutasyon bilgileri ile genişletilmiştir. Buna ek olarak, DGIdb'den elde edilen ilaç–gen etkileşim bilgileri de entegre edilmiştir. Böylelikle farmakolojik ve genetik parametreleri bir arada barındıran çok boyutlu bir veri seti oluşturulmuştur.

Deneysel çalışmaların tümü Python programlama dili kullanılarak yürütülmüştür. Model geliştirme aşamalarında XGBoost, CatBoost ve scikit-learn kütüphanelerinden yararlanılmıştır. Veri seti, %80 eğitim ve %20 test olacak şekilde ikiye bölünmüştür.

4.2.2. İlaç–Gen etkileşimlerinin incelenmesi

Model performansını artırmak ve aşırı öğrenmeyi (overfitting) önlemek amacıyla veri seti üzerinde özellik seçimi uygulanmıştır. İlk aşamada, eğitim veri setinde sıfır varyansa sahip olan ve tüm örneklerde aynı değeri alan değişkenler VarianceThreshold yöntemi kullanılarak veri setinden çıkarılmıştır. Daha sonra kalan değişkenler için f_regression testi uygulanarak bağımlı değişken olan Z_SCORE ile istatistiksel olarak anlamlı ilişkiye sahip özellikler belirlenmiştir. Bu kapsamda p-değeri 0.05'ten küçük olan değişkenler anlamlı kabul edilerek modelleme sürecine

dahil edilmiştir. Anlamlı değişken bulunmaması durumunda ise model tüm mevcut değişkenler kullanılarak eğitilmeye devam etmiştir.

Seçilen 11 özellik kullanılarak XGBoost, CatBoost, LightGBM, Random Forest ve Extra Trees olmak üzere beş farklı regresyon modeli eğitilmiştir. Bu ağaç tabanlı algoritmalar, yüksek genelleme yetenekleri, eksik verilerle çalışabilme kapasiteleri ve yüksek boyutlu biyolojik veri setlerinde yorumlanabilir sonuçlar sunmaları nedeniyle tercih edilmiştir. Modellerin performansını artırmak amacıyla hiperparametre optimizasyonu gerçekleştirilmiş ve her model için en uygun parametreler RandomizedSearchCV yöntemi ile belirlenmiştir. Bu süreçte n_estimators, max_depth, learning_rate ve diğer ilgili hiperparametreler sistematik olarak test edilerek en uygun değerler seçilmiştir. Böylece modeller gereksiz özelliklerden arındırılmış, daha düşük karmaşıklıkta ve yüksek doğrulukta olacak şekilde eğitilmiştir.

Önerilen analiz çerçevesinin daha kapsamlı değerlendirilmesi amacıyla, başlangıçta GDSC veri seti için geliştirilen yöntem farklı bir farmakogenomik bağlamda da uygulanmıştır. Bu kapsamda Cancer Therapeutics Response Portal (CTRP) veri tabanındaki ilaç yanıt verileri ile Cancer Cell Line Encyclopedia (CCLE) veri tabanındaki çoklu omik veriler entegre edilmiştir. GDSC veri setinde ilaç duyarlılığı genellikle log(IC50) değerleri ile ifade edilirken, CTRP veri setinde ilaç yanıtı AUC (dose–response eğrisi altındaki alan) ile raporlanmaktadır. CTRP ilaç–hücre hattı yanıt verileri, DepMap kimlikleri kullanılarak CCLE genomik verileri ile eşleştirilmiştir. CCLE'den elde edilen kopya sayısı değişimleri ve somatik mutasyon verileri CTRP ilaç yanıt verileri ile birleştirilmiş ve GDSC veri seti için uygulanan özellik mühendisliği, kodlama ve özellik seçimi adımları aynen uygulanarak analiz edilmiştir.

4.2.2.1. Etkileşim sıklığı analizi

Analizlerin ardından, en çok gen ile etkileşime giren ilaçlar ile en fazla ilaçla ilişkili olan genler ortaya çıkarılmıştır. Sonuçlara göre, 273 numaralı ilaç (CUDC-101) 1089 etkileşim ile listenin başında yer almaktadır. Bu ilacı 158 numaralı ilaç (PF-562271) 520 etkileşim ile ve 1401 numaralı ilaç (AZD5438) 359 etkileşim ile

takip etmektedir. Genlerle en yoğun etkileşim gösteren ilk beş ilaç Tablo 5'te sunulmuştur.

Tablo 5. En fazla gen etkileşimine sahip ilaçlar ve farmakolojik özellikleri

Sıra	İlaç Kimliği (Drug_ID)	Etkileşim Sayısı	İlaç Adı	İlaç Özelliği
1	273	1089	CUDC-101	Çoklu hedefli inhibitör (HDAC, EGFR ve HER2 inhibitörü)
2	158	520	PF-562271	FAK inhibitörü (Focal Adhesion Kinase inhibitörü)
3	1401	359	AZD5438	Siklin bağımlı kinaz inhibitörü (CDK1/2/9 inhibitörü)
4	1490	279	SN-38	Topoizomeraz I inhibitörü (irinotekanın aktif metaboliti)
5	1049	272	PD173074	FGFR inhibitörü (Fibroblast Growth Factor Receptor inhibitörü)

Genler açısından değerlendirildiğinde ise COSG115384 (TP53) toplam 868 etkileşim ile en çok ilaçla ilişkilendirilen gen olarak öne çıkmaktadır. Onu COSG73399 (EGFR) 670 etkileşim ve COSG89395 (ERBB2) 495 etkileşim ile izlemektedir. Tablo 6'da en fazla ilaç etkileşimine sahip ilk beş gen gösterilmiştir. Bu sonuçlar, ilaçlar ile genler arasındaki etkileşimlerin kanser tedavisinde önemli terapötik hedefler olabileceğini ortaya koymaktadır.

Tablo 6 . İlaçlarla en yoğun etkileşime giren genler ve fonksiyonları

Sıra	COSMIC Gen Kimliği	Etkileşim Sayısı	Gen Adı	Genin Kısa Açıklaması
1	COSG115384	868	TP53	DNA hasarına yanıt veren ve hücre döngüsünü kontrol eden önemli bir tümör baskılayıcı gen
2	COSG73399	670	EGFR	Hücre proliferasyonu ve sinyal iletiminde rol oynayan reseptör tirozin kinaz
3	COSG89395	495	ERBB2	HER2 olarak da bilinen, hücre büyümesi ve farklılaşmasını düzenleyen reseptör tirozin kinaz
4	COSG112841	357	HDAC9	Histon deasetilaz ailesine ait olup epigenetik gen düzenlenmesinde görev alır
5	COSG114519	348	FGFR1	Hücre büyümesi, farklılaşması ve anjiyogenezde rol oynayan fibroblast büyüme faktörü reseptörü

4.2.2.2. İlaç duyarlılık analizi

Çalışmada ilaçların etkinliğini değerlendirmek amacıyla Z-skorları temel alınmıştır. Literatürde, negatif Z-skorlarının hücre büyümesini güçlü şekilde baskıladığı, pozitif Z-skorlarının ise düşük etkinliğe karşılık geldiği bildirilmektedir (Yang ve ark., 2013). En fazla ilaç etkileşimine sahip ilk beş gen için kapsamlı bir ilaç duyarlılık analizi yapılmış; her gen için Z-skorlarına dayalı yanıt profilleri oluşturularak duyarlılık paternleri belirlenmiştir. Sonuçlar Şekil 4'te gösterilmiştir. Analiz sonuçları:

✓ Z-score değerleri kullanılarak her gen için ilaç yanıt profilleri oluşturulmuştur.

✓ TP53 (COSG115384) geni birçok ilaca karşı geniş bir yanıt spektrumu göstermiştir. SN-38 (İlaç 1490) ilacına karşı yüksek duyarlılık saptanmıştır (Z-score: −1.5).

✓ EGFR (COSG73399) geni genel olarak negatif Z-score değerleri göstermiş ve Palbociclib (İlaç 87) için Z-score −1.0 bulunmuştur.

✓ ERBB2 (COSG89395) geni ile Palbociclib (İlaç 87) arasında ilaç duyarlılığı gözlenmiştir (Z-score: −0.8).

✓ HDAC9 (COSG112841) geni tüm ilaçlara karşı pozitif Z-score değerleri göstermiştir; bu durum ilaç duyarlılığı ile ilişkili olabileceğini düşündürmektedir.

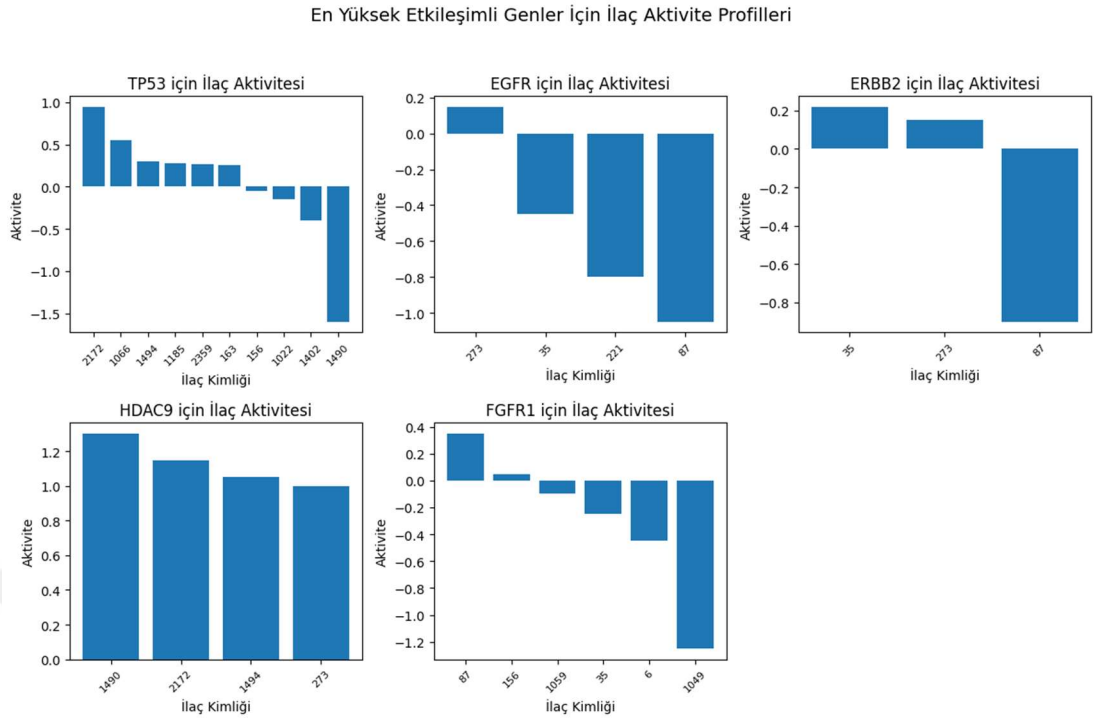
✓ FGFR1 (COSG114519) geni değişken ilaç yanıtı göstermiş ve Mirin (İlaç 1048) ilacına karşı yüksek duyarlılık saptanmıştır (Z-score: −1.2).

✓ Bulgular, gen-ilaç etkileşimlerinin karmaşık ve çok faktörlü olduğunu göstermektedir.

✓ EGFR ve ERBB2 gibi genler potansiyel terapötik hedefler olarak öne çıkmaktadır.

✓ HDAC9 geninin ise ilaç direnci gelişiminde rol oynayabileceği düşünülmektedir.

Şekil 4'ün sonuçları, her bir genin ilaçlara karşı kendine özgü bir yanıt verdiğini ve gen–ilaç ilişkilerinin çok boyutlu, karmaşık bir yapıya sahip olduğunu göstermektedir. Negatif Z-skorları baskın olan EGFR (COSG73399) ve ERBB2 (COSG89395) genleri, kanser tedavisinde umut vadeden hedefler olarak öne çıkmaktadır. Buna karşın, test edilen tüm ilaçlara karşı pozitif Z-skoru sergileyen HDAC9 (COSG112841) geninin ilaç duyarlılığının oluşumunda önemli rol üstlenebileceği düşünülmektedir. Dahası, Palbociclib (İlaç 87)'nin birçok gen üzerinde gözlenen güçlü negatif etkisi, bu bileşiğin geniş spektrumlu antikanser etkinlik potansiyeline sahip olabileceğini göstermektedir.



Şekil 4. En fazla etkileşime giren 5 genin ilaç yanıt profilleri (Z-skoru değerleri).

4.2.3. Kümeleme ve korelasyon analizleri

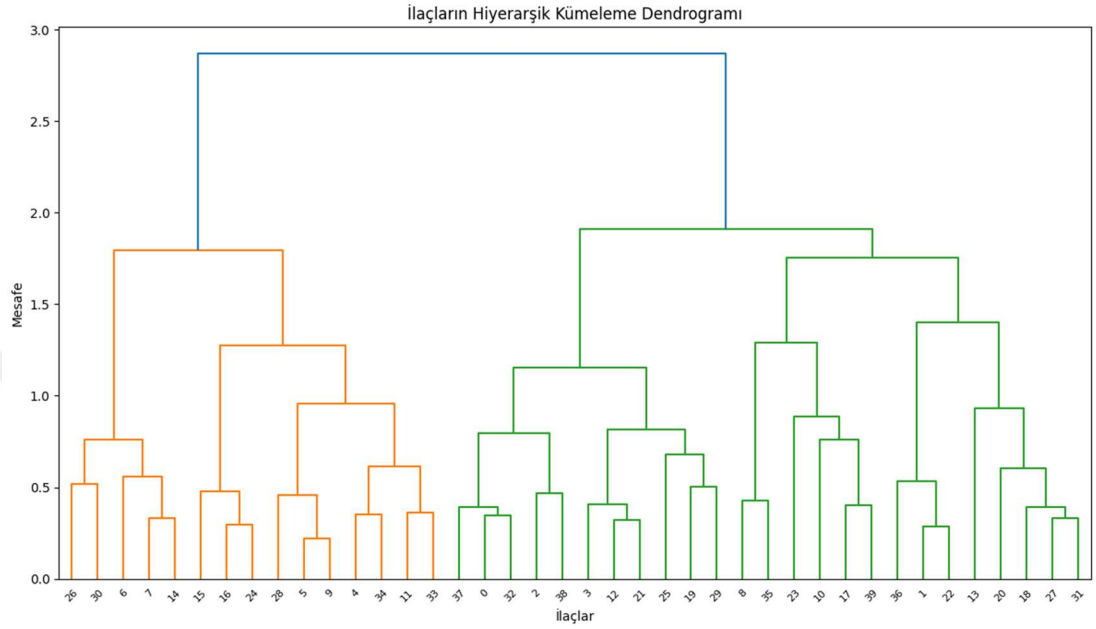
4.2.3.1 Hiyerarşik kümeleme bulguları

Gen yanıt profillerine göre ilaçların sınıflandırılması amacıyla Ward hiyerarşik kümeleme yöntemi uygulanmıştır. Ortaya çıkan dendrogramda, ilaçlar arasındaki benzerlikler mesafe değerleri üzerinden görselleştirilmiş ve benzer genetik yanıt gösteren ilaçlar aynı gruplar içinde kümelenmiştir (Şekil 5).

Analiz sonuçlarına göre, dendrogramın sol tarafında turuncu ile işaretlenen düğüm diğer ilaç gruplarından belirgin biçimde ayrılmakta ve farklı bir etkinlik profiline işaret etmektedir. Bu kümeye dâhil olan ilaçlar, birbirine yakın mesafelerde gruplanmış ve benzer gen yanıt paternleri sergilemiştir. Bu durum, benzer etki mekanizmasına veya farmakodinamik özelliklere sahip ilaçların analizde doğru biçimde aynı grupta toplandığını göstermektedir.

Bu bulgular, ilaçların genetik aktivite profillerine göre düzenli bir şekilde kümelenebileceğini ve bu yaklaşımın potansiyel benzer biyolojik etkilerin veya kombinasyon tedavi adaylarının belirlenmesine yardımcı olabileceğini

göstermektedir. Özellikle aynı kümede yer alan ilaçların ortak biyolojik yolakları hedefleme veya birbirlerinin tedavi etkinliğini artırma ihtimali dikkat çekmektedir.



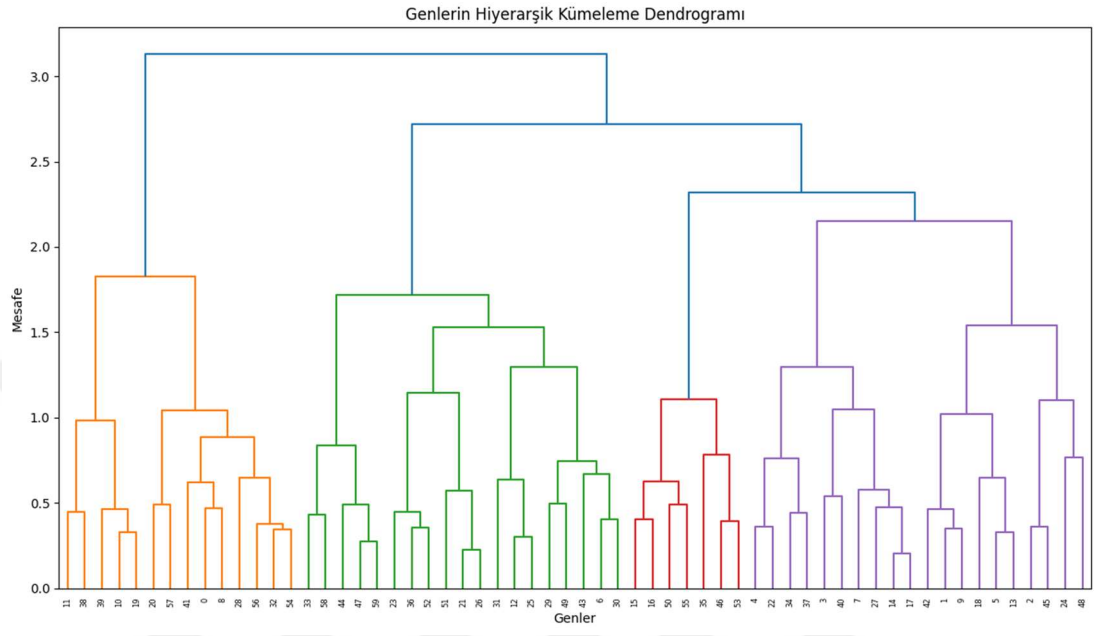
Şekil 5. Gen yanıt profillerine göre ilaçların hiyerarşik kümeleme dendrogramı.

Mesafe ekseni, ilaçlar arasındaki genetik aktivite benzerliklerini göstermektedir. Sol tarafta turuncu ile işaretlenen küme, diğer gruplardan açıkça ayrışmakta ve özgün bir aktivite paterni sergilemektedir.

Ward yöntemi kullanılarak genlerin farklı ilaçlara karşı etkinlik yanıtları analiz edilmiş ve hiyerarşik kümeleme dendrogramı oluşturulmuştur. Dendrogramda genler arasındaki mesafe benzerlikleri, aynı ilaç yanıt profillerine sahip olanların bir araya toplanmasını sağlamıştır. Bu kümeler Şekil 6'da görselleştirilmiştir.

Analiz sonuçları, dendrogramın sol tarafında mavi ile işaretlenen kümenin diğer gen gruplarından ayrıldığını ve benzersiz bir ilaç yanıt profili ortaya koyduğunu göstermiştir. Bu kümedeki genler birbirine yakın konumlanmış ve benzer etkinlik profilleri sergilemiştir. Buna karşılık, turuncu ile gösterilen kümeler daha geniş dağılımlara sahip olsalar da kendi içlerinde tutarlı yanıt paternlerini korumuşlardır.
Sonuçlar, genlerin ilaçlara verdikleri yanıtların sistematik olarak kümelendirilebileceğini ortaya koymaktadır. Aynı kümede bulunan genlerin genellikle ortak biyolojik işlevlere katıldığı ya da benzer direnç/duyarlılık

mekanizmaları sergilediği anlaşılmaktadır. Bu bulgular, ilaç yanıtlarını öngörmeye yönelik güvenilir makine öğrenmesi modelleri geliştirilmesi ve daha etkili hedefe yönelik tedavilerin tasarlanması için değerli bir temel sağlamaktadır.



Şekil 6. İlaç yanıt profillerine göre genlerin hiyerarşik kümeleme dendrogramı.

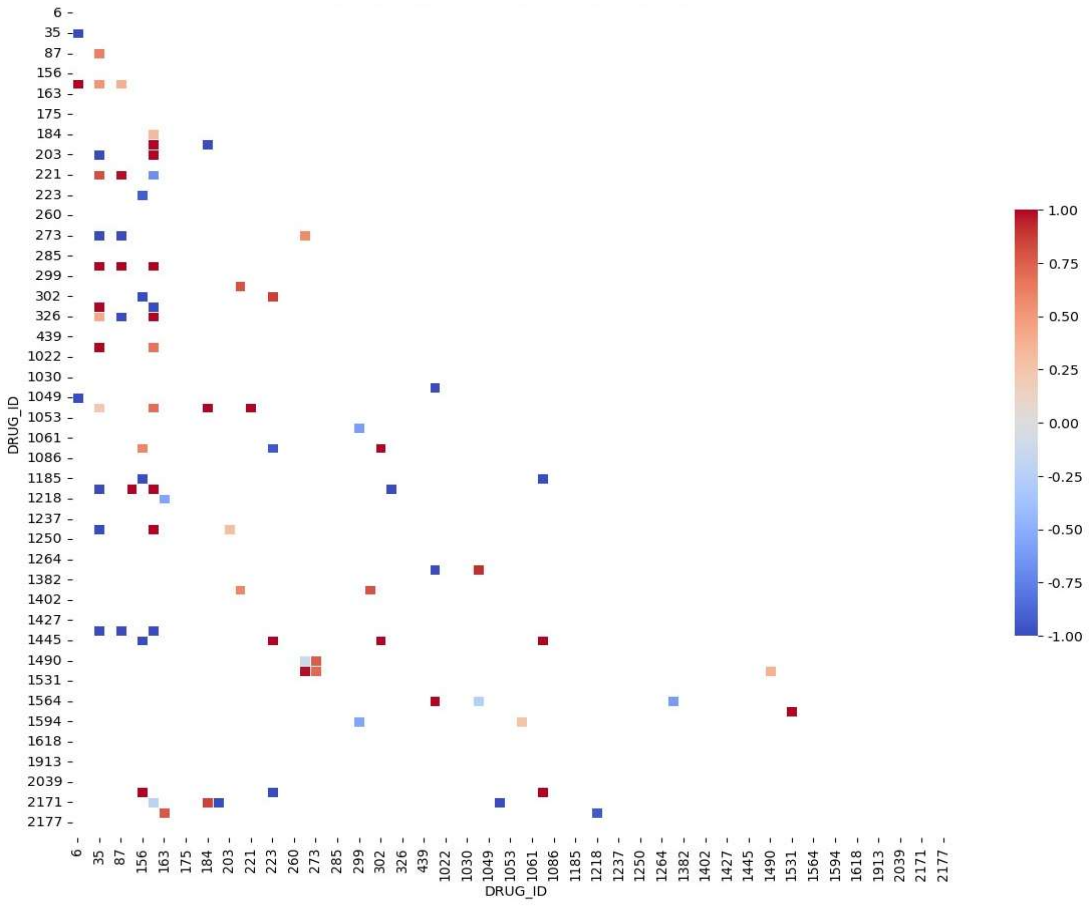
Mesafe ekseni, genler arasında ilaç etkinliği açısından benzerlik düzeyini göstermektedir. Sol tarafta mavi ile işaretlenen küme diğer gen gruplarından açık biçimde ayrılmakta ve farklı bir ilaç yanıt profili ortaya koymaktadır. Turuncu ile işaretlenen kümeler ise kendi içinde benzer yanıt gösteren gen topluluklarını yansıtmaktadır.

4.2.3.2. Korelasyon analizine ilişkin bulgular

İlaç–gen korelasyon analizleri, ilaçların etkinliğine dayalı ilişkilerin daha ayrıntılı bir biçimde anlaşılmasına imkân tanımaktadır. Bu analizler, biyolojik etkileşim ağlarının keşfine katkı sağlamakta ve farmakolojik benzerliklerin yanı sıra ilaçlar ile genlerin farklı etkinlik profillerini ortaya koymaktadır.

Şekil 7' de yer alan ilaç korelasyon ısı haritası, etkinlik profillerindeki benzerlikleri korelasyon katsayıları kullanılarak renkler üzerinden göstermektedir. Renk skalasında kırmızı tonlar pozitif korelasyonu (benzer ilaç profilleri), mavi

tonlar ise negatif korelasyonu (zıt ilaç profilleri) temsil etmektedir. Korelasyon katsayılarının -1 ile +1 arasında değiştiği bu görselleştirme, ilaçların etkinlik açısından birbirine ne ölçüde benzer veya farklı olduğunu açıkça ortaya koymaktadır. Sonuçlar incelendiğinde, bazı ilaç gruplarının güçlü pozitif korelasyon sergilediği ve benzer farmakodinamik özellikler taşıyabileceği görülmüştür. Kırmızı bölgeler benzer etkinlik gösteren ilaç gruplarını işaret ederken; mavi bölgeler ise farklı etki mekanizmalarına sahip ilaçlar arasında ters yönlü etkinlik ilişkilerini göstermektedir.



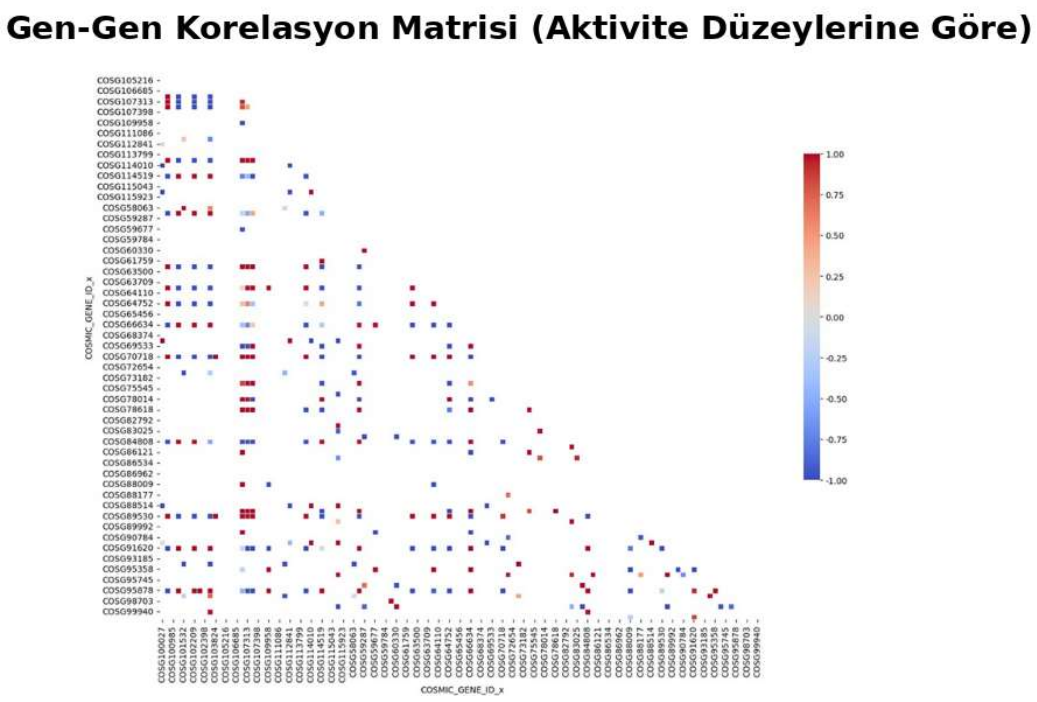
Şekil 7. İlaçların etkinlik korelasyonunu gösteren ısı haritası.

Renk skalasında kırmızı tonlar pozitif korelasyonları, mavi tonlar ise negatif korelasyonları ifade etmektedir. Özellikle yoğun kırmızı bölgeler, ilaçların benzer etkinlik profilleri sergilediğini ve farmakolojik etkilerinin örtüştüğünü göstermektedir.

Gen korelasyon analizine ait bulgular Şekil 8' de sunulmuştur. Bu analizde

genlerin farklı ilaçlara karşı gösterdikleri etkinlik yanıtları arasındaki ilişkiler değerlendirilmiş, benzer yanıt profillerine sahip gen kümeleri ortaya çıkarılmıştır. İlaç korelasyon ısı haritasında olduğu gibi, renk kodlaması korelasyon katsayıları temelinde yapılmıştır. Kırmızı tonlar benzer yanıt profillerini gösteren pozitif korelasyonu, mavi tonlar ise ters yanıt profillerini işaret eden negatif korelasyonu temsil etmektedir.

Sonuçlar, güçlü pozitif korelasyon gösteren gen çiftlerinin ortak düzenleyici ağlarda veya benzer biyolojik süreçlerde yer alma ihtimalini ortaya koymuştur. Kırmızı bölgeler aynı etki paternine sahip gen çiftlerini gösterirken, negatif korelasyon sergileyen genler karşıt biyolojik süreçlere veya zıt ilaç yanıt mekanizmalarına işaret etmektedir. Bu bulgular, potansiyel ilaç hedeflerinin belirlenmesi, biyobelirteç adaylarının tanımlanması ve kombinasyon tedavi stratejilerinin geliştirilmesi açısından önemli katkılar sağlamaktadır.



Şekil 8. Genler arası aktivite korelasyonunu gösteren ısı haritası

Renk skalasında kırmızı tonlar genler arasında pozitif korelasyonu, mavi tonlar ise negatif korelasyonu göstermektedir. Özellikle yoğun kırmızı alanlar, genlerin benzer aktivite yanıtları verdiğini ve aynı biyolojik süreçlerde görev alabileceklerini

ortaya koymaktadır.

4.2.4. Makine öğrenmesi modellerinin performans analizi

Bu bölümde çalışmada kullanılan makine öğrenmesi modellerinin başarıları karşılaştırılmıştır. Tablo 7'de LightGBM Regressor, XGBoost Regressor, CatBoost Regressor, Extra Trees Regressor ve Random Forest Regressor algoritmalarının performans değerlendirmeleri ile en iyi hiperparametre sonuçları sunulmuştur.

Elde edilen karşılaştırmalara göre, eğitim ve test süreçlerinde en yüksek R^2 değeri LightGBM Regressor tarafından sağlanmıştır. Bu nedenle, ilaç–gen etkileşimlerinde Z_SCORE tahmini için en uygun model LightGBM Regressor olarak belirlenmiştir. Tablo 7, her bir modelin en iyi hiperparametre değerlerini, R^2 skorlarını ve ilgili performans metriklerini göstermektedir.

Tablo 7. Eğitilmiş Modellerin Hiperparametre Ayarları ve Çapraz Doğrulama R^2 Skorları

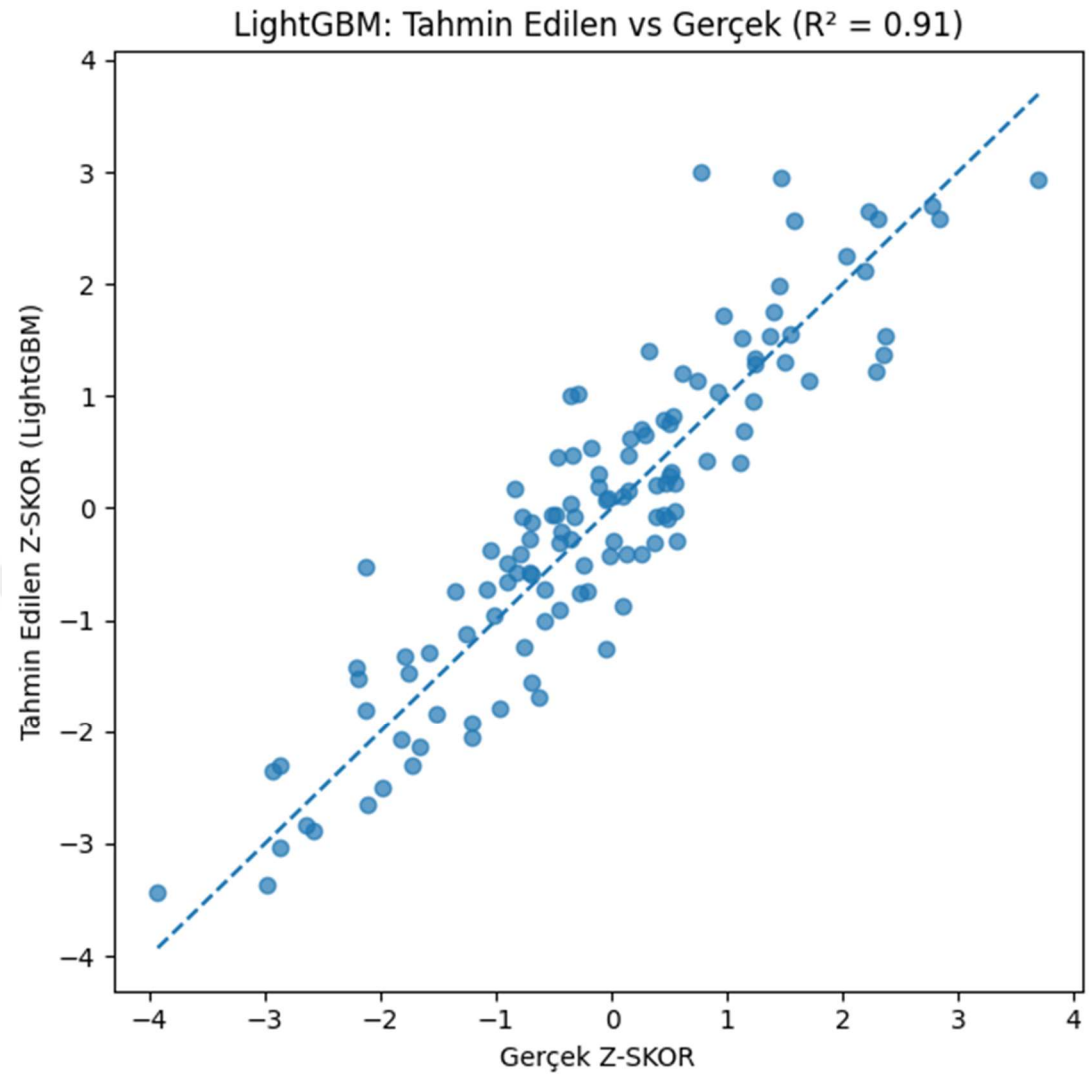
Model	En İyi Hiperparametreler	R^2	MSE	MAE	AUC
LightGBM Regressor	learning_rate = 0.229; num_leaves = 40; n_estimators = 409	0.9120	0.1371	0.1576	0.9680
XGBoost Regressor	colsample_bytree = 0.77; learning_rate = 0.072; max_depth = 6; n_estimators = 409	0.9110	0.1385	0.1816	0.9674
CatBoost Regressor	depth = 6; l2_leaf_reg = 8; learning_rate = 0.189	0.9105	0.1394	0.1745	0.9671
Extra Trees Regressor	max_depth = 17; max_features = 0.421; n_estimators = 180	0.9082	0.1430	0.1642	0.9666
Random Forest Regressor	max_depth = 17; max_features = 0.421; n_estimators = 180	0.9066	0.1455	0.1755	0.9675

Tablo 7 sonuçlarına göre LightGBM Regressor en yüksek başarıyı göstermiştir. Model $R^2 = 0{,}9120$ değeriyle öne çıkarken, aynı zamanda en düşük MSE (0,1371) ve MAE (0,1576) değerlerini üretmiştir. XGBoost ve CatBoost modelleri LightGBM'e çok yakın sonuçlar elde etmiş; $R^2 \approx 0{,}9110$ ve 0,9105, MSE ≈ 0,1385 ve 0,1394, MAE ≈ 0,1816 ve 0,1745, AUC ≈ 0,9674 ve 0,9671 değerleriyle güçlü performans

sergilemiştir. Extra Trees ve Random Forest modelleri hata metrikleri açısından bir miktar geride kalsa da, AUC değerlerinin 0,9666–0,9675 aralığında yer alması bu algoritmaların ikili sınıflandırmada hâlâ yüksek ayırt edici güce sahip olduklarını göstermektedir. Genel olarak tüm ağaç tabanlı modeller başarılı olsa da, LightGBM Regressor tahmin doğruluğu açısından bir adım öndedir.

LightGBM'in performansı ayrıntılı biçimde incelendiğinde, Şekil 9'da modelin gerçek ve tahmin edilen Z_SCORE değerleri arasındaki ilişki görülmektedir. Kırmızı kesikli çizgi ideal tahmin doğrusunu (y = x) temsil ederken, veri noktalarının bu çizgiye yakın şekilde kümelenmesi modelin güçlü bir açıklayıcılığa sahip olduğunu ($R^2 = 0,91$) göstermektedir. Özellikle orta değer aralığında noktaların yoğun şekilde toplanması, LightGBM'in en sık görülen gözlemler için tutarlı tahminler yaptığını; uç değerlerdeki sınırlı sapmalar ise modelin genelleme kapasitesinin korunduğunu göstermektedir.

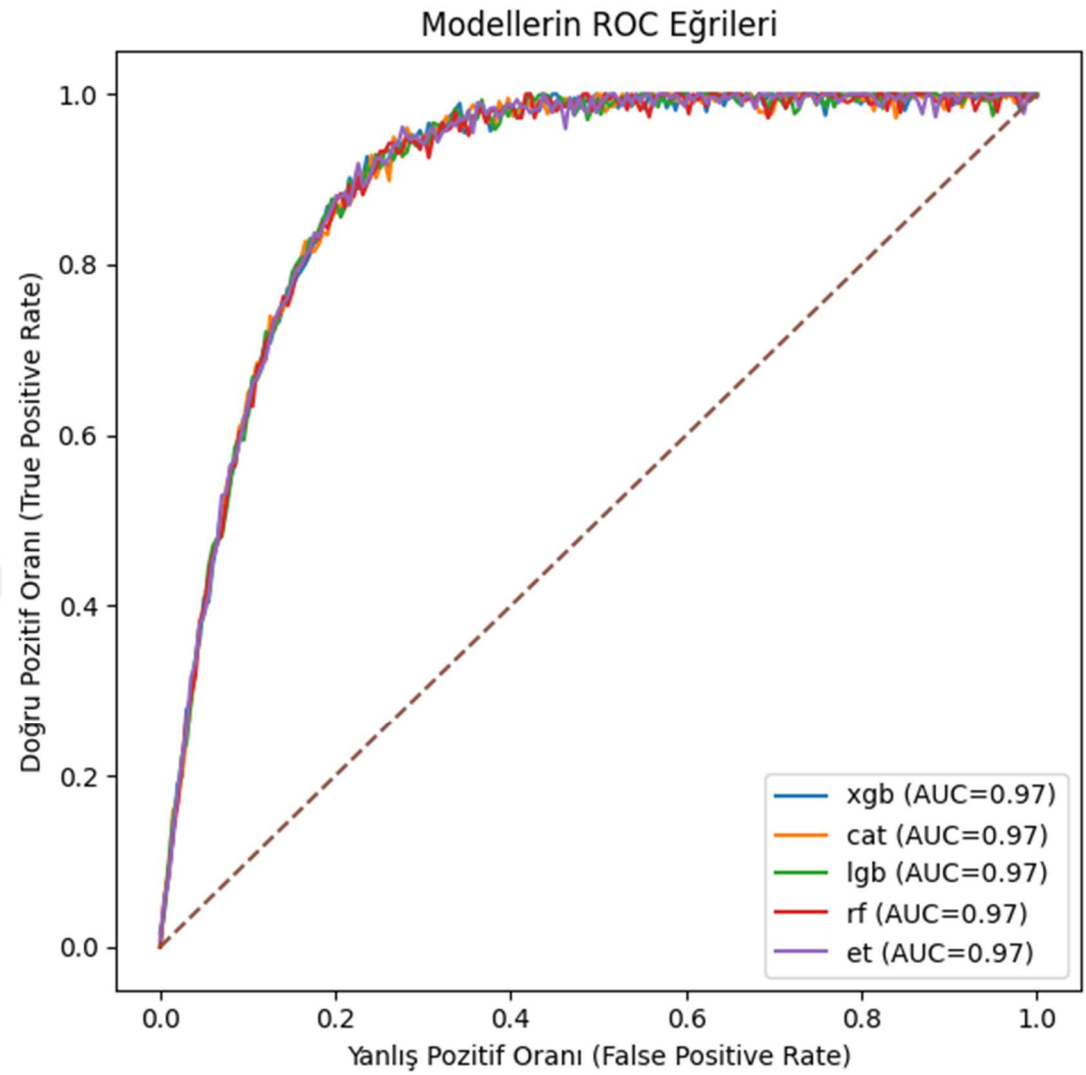
Sonuç olarak, LightGBM Regressor mevcut veri seti üzerinde en iyi tahmin performansına ulaşmıştır. Başka bir ifadeyle, Şekil 9, LightGBM'in mevcut dağılım koşullarında optimuma yakın öngörü gücü sunduğunu ortaya koymaktadır.



Şekil 9. En iyi performans gösteren modelin gerçek ve tahmin edilen değerleri için R^2 dağılım grafiği.

İkili sınıflandırma performanslarını karşılaştırmak üzere, Şekil 10' da XGBoost, CatBoost, LightGBM, Random Forest ve Extra Trees algoritmalarına ait ROC eğrileri gösterilmiştir. Elde edilen eğriler büyük ölçüde birbirleriyle örtüşmekte ve düşük yanlış pozitif oranlarında bile %95'in üzerinde doğru pozitif oranı sağlamaktadır. AUC ≈ 0,97 değerleri, modellerin hem duyarlılık hem de özgüllük açısından yüksek başarıya ulaştığını ortaya koymaktadır.

Sonuç olarak, veri seti yüksek ve düşük Z_SCORE sınıfları arasında güçlü bir ayrışma sağlamaktadır. Bu bulgular, uygulanan ön işleme ve özellik seçimi yöntemlerinin hem sürekli tahmin hem de ikili sınıflandırma görevlerinde yüksek düzeyde genelleme kapasitesi sunduğunu teyit etmektedir (Şekil 10).



Şekil 10. Ağaç tabanlı algoritmalara ait ROC eğrileri.

4.2.5. Ablasyon analizi

Çalışmada, özelliklerin öğrenmeye katkısını incelemek amacıyla dört farklı ablasyon senaryosu tasarlanmıştır. Bu sayede, önerilen ön işleme ve özellik seçimi adımlarının model başarısına katkısı net biçimde ölçülmüştür. Şekil 11, işlem hattının her bileşeninin metodolojiye ve elde edilen sonuçlara etkisini özetlemektedir.

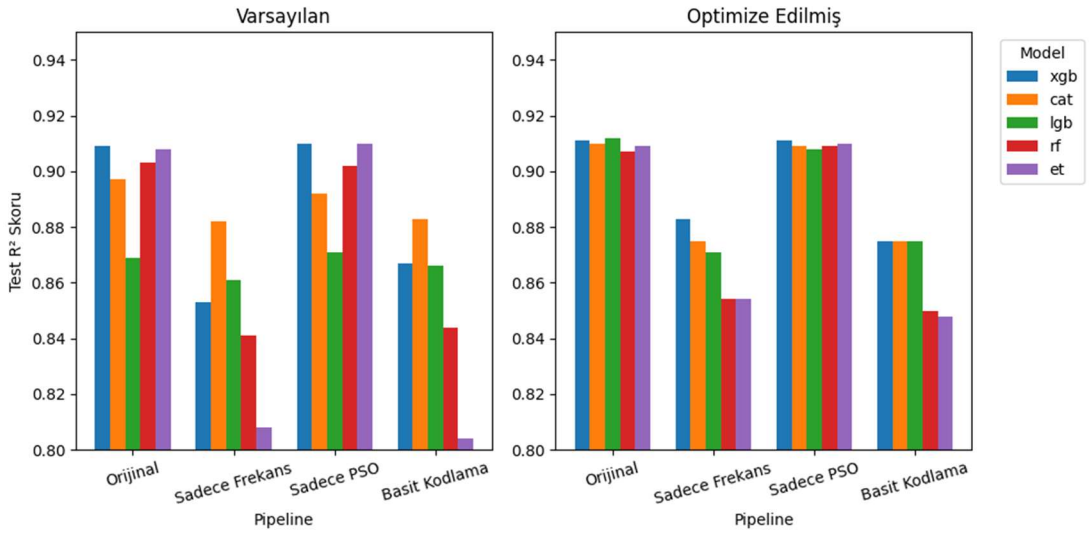
Her iş akışında XGBoost, CatBoost, LightGBM, Random Forest ve Extra Trees algoritmaları önce varsayılan hiperparametrelerle, ardından RandomizedSearchCV yöntemiyle optimize edilmiş hiperparametrelerle eğitilmiş; bağımsız test setinde R^2, MSE, MAE ve AUC performans metrikleri üzerinden değerlendirilmiştir.

Orijinal işlem hattı, düşük kardinalite için One-Hot Encoding, orta kardinalite için Target Encoding, yüksek kardinalite için Ordinal Encoding aşamalarını; ardından VarianceThreshold, f_regression ve sabit PSO tabanlı özellik seçimini içermektedir.

Freg only yaklaşımında sadece VarianceThreshold ve f_regression adımları uygulanmıştır.

PSO only senaryosunda, VarianceThreshold sonrası yalnızca PSO ile belirlenen 11 özellik kullanılmıştır.

Simple Encode iş akışında tüm kategorik değişkenler yalnızca Label Encoding yöntemiyle dönüştürülmüş, ek bir özellik seçimi yapılmamıştır. Bu farklı senaryolardan elde edilen R^2 skorları Şekil 11'de gösterilmiştir.



Şekil 11. Ablasyon Analizi: Varsayılan ve optimize edilmiş modellerin işlem hattına göre karşılaştırmalı R^2 sonuçları.

Orijinal işlem hattında en yüksek performans, optimize edilmiş LightGBM modeliyle elde edilmiş (R^2 = 0,9120, MSE = 0,1371, MAE = 0,1576, AUC = 0,9680). Bu modeli XGBoost (R^2 = 0,9110) ve CatBoost (R^2 = 0,9105) çok yakın değerlerle takip etmiştir.

Freg-only yaklaşımında, optimize edilmiş XGBoost belirgin şekilde daha düşük performans göstermiştir (R^2 = 0,8824, MSE = 0,1833, MAE = 0,2366, AUC = 0,9614). Bu sonuç, orijinal pipeline'a göre kayda değer bir düşüşe işaret etmektedir.

PSO-only yaklaşımında, seçilen 11 özellik üzerinden yapılan eğitimler orijinal

pipeline'a yakın sonuçlar vermiştir. Örneğin optimize edilmiş XGBoost $R^2 = 0{,}9110$, Extra Trees $R^2 = 0{,}9101$ ve LightGBM $R^2 = 0{,}9081$ değerlerine ulaşmıştır.

En basit dönüşüm stratejisi olan Simple Encode senaryosunda ise, optimize edilmiş XGBoost modeli $R^2 = 0{,}8746$, MSE = 0,1953, MAE = 0,2557, AUC = 0,9602 performansı göstermiş, diğer algoritmalar da benzer şekilde geride kalmıştır. Sonuç olarak, bu bulgular çok aşamalı özellik seçimi ve kapsamlı kodlama stratejilerinin biyoinformatik etkileşim verilerinde hem doğruluk hem de genellenebilirlik için kritik rol oynadığını kanıtlamaktadır.

4.2.6. Bağımsız CTRP–CCLE Farmakogenomik Verisi Üzerinde Değerlendirme

Önerilen analiz hattının yaygın olarak kullanılan farmakogenomik veri setlerinin ötesinde genellenebilirliğini değerlendirmek amacıyla, tüm modelleme çerçevesi Cancer Therapeutics Response Portal (CTRP) veri setinden elde edilen ilaç yanıt profilleri ile Cancer Cell Line Encyclopedia (CCLE) veri setinden elde edilen çoklu-omik verilerin entegrasyonu kullanılarak uygulanmıştır.

GDSC gibi veri setlerinde ilaç duyarlılığı genellikle log-dönüştürülmüş IC50 değerleri ile temsil edilirken, CTRP veri setinde ilaç yanıtı doğrudan doz–yanıt eğrisi altındaki alan (AUC) olarak raporlanmaktadır. Bu nedenle tahmin görevi, ilaç duyarlılığı SENS = –AUC şeklinde tanımlanarak oluşturulmuştur. Bu yaklaşım, CTRP veri setinin orijinal yanıt ölçeğinin korunmasını sağlamaktadır.

Tüm analiz hattı; genomik özellik oluşturma, bilgi sızıntısını önleyen kategorik kodlama stratejileri, meta-sezgisel özellik seçimi ve gradyan artırmalı regresyon modellerini içerecek şekilde veri setine özgü ek ayarlamalar yapılmadan uygulanmıştır. CCLE'den elde edilen genomik bilgiler hem toplu tanımlayıcılar hem de geniş özet istatistikler kullanılarak özetlenmiş; ilaç ve dokuya ait meta veriler ise yalnızca eğitim verisi üzerinden öğrenilen kontrollü kodlama yöntemleri ile modele dahil edilmiştir.

Genişletilmiş sayısal özellik havuzundan Parçacık Sürü Optimizasyonu (Particle Swarm Optimization, PSO) yöntemi ile seçilen bilgilendirici bir özellik alt kümesinden yararlanılarak, LightGBM modeli $R^2 = 0.74$ değeri ile umut verici bir tahmin performansı göstermiştir.

Veri işleme hattı algoritması, CTRP ve CCLE veri setlerinin entegrasyonu için otomatik olarak 13 özellik seçmiştir. Bu özellikler; ilaç tanımlayıcıları (DRUG_NAME, TCGA_DESC), çeşitli Kopya Sayısı Değişimi (Copy Number Variation, CNV) metrikleri (cnv_amp_minus_del, cnv_burden, cnv_mean_wide, cnv_q90, cnv_q90_q10_range, log1p_cnv_del_lt_0_90), mutasyon profilleri (mut_sum_wide, mut_nonzero_genes) ve transkriptomik yapının istatistiksel tanımlayıcılarını (tcga_mean, dt_mean, dt_n) kapsamaktadır. Elde edilen modelin performans sonuçları Tablo 5'te özetlenmiştir.

Tablo 8. AUC tabanlı ilaç duyarlılığına göre CTRP–CCLE veri entegrasyonunda model performansı

Hedef Değişken	R^2	MSE	MAE	Özellik Kümesi	Nihai Özellik Sayısı
AUC	0.7376	1.7672	0.9229	43	13

Yaklaşık 0.74'lük bir R^2 değeri, önerilen çerçevenin farklı kanser hücre hatları ve bileşikler arasında mutlak ilaç duyarlılığındaki değişkenliğin önemli bir kısmını açıklayabildiğini göstermektedir. İlaç duyarlılığı tahmini için GDSC veri setini kullanan son çalışmaların karşılaştırılması Tablo 9'da gösterilmiştir.

Tablo 9. İlaç duyarlılığı tahmini için GDSC veri setini kullanan son çalışmaların karşılaştırılması.

Çalışma	Veri Seti	Algoritma	Tahmin Değişkeni	R^2
Chawla ve ark. (2022) [50]	GDSC / CCLE	Derin Öğrenme	LN_IC50	0.77
Schlüter ve Schönhuth (2025) [51]	GDSC / CMP	Ridge Regresyon	IC50	0.87
Li ve ark. (2024) [52]	GDSC / CCLE	Derin Öğrenme	LN_IC50	0.88
Yu ve Fan ve ark. (2025) [53]	GDSC	Derin Öğrenme	LN_IC50	0.75
Önerilen Model	GDSC / COSMIC / DGIdb	LightGBM (Makine Öğrenmesi)	LN_IC50 / ZScore	0.91

Dikkat çekici olarak, bu performans nispeten küçük bir nihai özellik kümesi kullanılarak elde edilmiştir. Bu durum, PSO tabanlı özellik seçimi stratejisinin, model karmaşıklığını kontrol altında tutarken bilgilendirici genomik ve bağlamsal özellikleri belirlemede etkili olabileceğini göstermektedir.

5. TARTIŞMA VE SONUÇ

Deneysel çalışma iki temel aşamada yürütülmüştür. İlk aşamada, küçük hücreli akciğer kanseri hücre hatlarına ait mutasyon yüküne dayalı öznitelikler oluşturularak modellemeye uygun bir veri seti yapılandırılmıştır. Bu kapsamda mutasyon türü, sayısı, biyolojik etkisi ve varyant alel frekansı gibi özellikler sayısallaştırılmış ve her hücre hattı için tanımlayıcı değişkenler elde edilmiştir.

İkinci aşamada ise bu öznitelikler kullanılarak ilaç direncinin (IC50) tahmini amaçlanmıştır. Geliştirilen makine öğrenmesi modellerinin performansı RMSE, R^2 ve MAE metrikleri ile değerlendirilmiştir. Bu metriklerin birlikte kullanılması sayesinde model doğruluğu, hata büyüklüğü ve açıklayıcılık gücü kapsamlı biçimde analiz edilmiştir.

Sonuçlar, test verisi üzerinde en başarılı performansı LightGBM Regressor modelinin gösterdiğini ortaya koymuştur (R^2 = 0.9120; MSE = 0.1371; MAE = 0.1576). XGBoost ve CatBoost modelleri de benzer düzeyde yüksek performans sergilemiş, tüm gradyan artırma temelli yöntemlerin güçlü tahmin kapasitesine sahip olduğu görülmüştür. Extra Trees ve Random Forest modelleri hata metrikleri açısından bir miktar geride kalsa da, yüksek AUC değerleri bu algoritmaların ayırt edici gücünün sürdüğünü göstermektedir.

XGBoost modeline göre en belirleyici öznitelikler sırasıyla nonsense, frameshift, toplam mutasyon yükü, upstream ve start-loss mutasyon sayılarıdır. Bu mutasyon türlerinin artışı, belirli ilaçlara karşı hücresel duyarlılığın değişebileceğini göstermektedir. Bulgular, ilaç seçimi öncesinde genetik profilin ayrıntılı değerlendirilmesinin önemini ortaya koymaktadır.

Genel olarak, ağaç tabanlı topluluk yöntemleri ilaç duyarlılığı tahmininde yüksek doğruluk sağlamış; özellikle LightGBM modeli performans açısından öne çıkmıştır.

Bu iki çalışma birlikte değerlendirildiğinde, genetik mutasyonların ilaç duyarlılığı üzerindeki belirleyici rolü hem dar kapsamlı (SCLC odaklı) hem de geniş kapsamlı (çoklu farmakogenomik veri entegrasyonu) analizlerle ortaya konmuştur. İlk çalışmada, küçük hücreli akciğer kanserinde mutasyon yükü ve mutasyon türlerinin ilaç direncine etkisi incelenmiş; özellikle nonsense, frameshift ve toplam mutasyon yükünün ilaç duyarlılığı üzerinde önemli etkisi olduğu gösterilmiştir. Bu

çalışma, mutasyon yüküne dayalı özelliklerin IC50 tahmininde kullanılabileceğini ve makine öğrenmesi yöntemlerinin deneysel çalışmalara yön gösterebilecek ön bilgi sağlayabileceğini ortaya koymuştur.

İkinci çalışmada ise farmakogenomik yaklaşım genişletilerek gen–ilaç–mutasyon etkileşimleri çoklu veri kaynaklarının entegrasyonu ile sistematik olarak analiz edilmiştir. Bu sayede yalnızca mutasyon yükü değil, aynı zamanda gen–ilaç etkileşim yoğunluğu ve farmakogenomik ilişkilerin ilaç duyarlılığı üzerindeki etkisi ortaya konmuştur. Geliştirilen modelin daha yüksek doğruluk ve genellenebilirlik göstermesi, çoklu veri entegrasyonunun ilaç yanıt tahmininde önemli bir avantaj sağladığını göstermektedir. Her iki çalışmanın ortak sonucu, ilaç duyarlılığının tek bir biyolojik faktörle açıklanamayacağı, ancak mutasyon yükü, mutasyon türleri ve gen–ilaç etkileşimlerinin birlikte değerlendirilmesiyle daha doğru tahmin edilebileceğidir.

Bu yaklaşım, kişiselleştirilmiş kanser tedavisi için kritik öneme sahip olup, hangi hastada hangi ilacın daha etkili olabileceğine dair öngörü sunabilecek potansiyel bir karar destek çerçevesi sağlayabilir.

Her iki çalışmanın da hücre hattı verilerine dayanması, klinik doğrulama gereksinimini ortaya koymaktadır. Ayrıca, gen ekspresyonu, epigenetik düzenlemeler ve tümör mikroçevresi gibi ek biyolojik faktörlerin modele dahil edilmesi, ilaç duyarlılığı tahmininin doğruluğunu daha da artırabilir. Gelecek çalışmalar, multi-omik veri entegrasyonu ve klinik hasta verileri ile doğrulama üzerine odaklanarak bu yaklaşımın translasyonel değerini güçlendirebilir.

Literatür ile karşılaştırıldığında, çalışma, mutasyon profilleri + gen–ilaç etkileşim bilgisini birleştirerek ilaç duyarlılığını hem açıklayan (istatistiksel analiz) hem de tahmin eden (ML) bütüncül bir çerçeve sunuyor.“Genomik belirteçlerden ilaç yanıtı” geleneği Cancer Cell Line Encyclopedia çalışmaları, genomik özelliklerden ilaç duyarlılığına yönelik modellerin mümkün olduğunu erken dönemde göstermiştir. Bu çizgide genetik korelatif ilişkiler ve tahminleme temeli atılmıştır. Çalışma bu çizgiyi devam ettirip, mutasyon + etkileşim bilgisi ile kapsamı genişletiyor. Makine öğrenmesi yaklaşımlarının genel çerçevesinde, ilaç yanıtı tahmininde integratif yaklaşımlar; matriks faktörizasyonu, kernel ve ağ-temelli yöntemler gibi farklı sınıflara ayrılır. Bu alandaki derlemeler, “çoklu veri” entegrasyonunun performans

ve sağlamlık için kritik olduğunun altını çizer.

Bu çalışmanın katkısı: entegrasyonu yalnızca "daha çok özellik" için değil, gen–ilaç ilişkisini biyolojik anlamlı bir katman olarak eklemek için kullanmasıdır. Yorumlanabilirlik ve "biyolojik tutarlılık" açısından değerlendirmek gerekirse, güncel çalışmalarda yorumlanabilir makine öğrenmesi çerçeveleri (ör. açıklanabilir/yorumlanabilir stratejiler) öne çıkmakta; amaç sadece skor üretmek değil, hangi genlerin neden etkili olduğunu göstermek olmalıdır.

Bu çalışmanın feature importance + ablasyon yaklaşımı bu trendle uyumlu: "tahmin + açıklama" birlikte verilmektedir. Veri tabanı temelli yenilik ise Drug–Gene Interaction Database'i, gen–ilaç etkileşimlerini bir araya getiren, hipotez üretimini kolaylaştıran önemli bir kaynak olarak kullanılıyor olmasıdır.

Çalışmamızın farkı: DGIdb benzeri etkileşim bilgisini model girdisi/özelliği gibi kullanarak mutasyon tabanlı sinyali "ilaç hedeflenebilirliği" perspektifiyle zenginleştirmektir. Literatürde bilinen zorluk olan ilaç yanıt verilerinde tekrar deneylerde bile korelasyonun orta düzeyde kalabildiği; veri setleri arası uyumun sınırlı olabildiği raporlanmıştır. Bu nedenle bağımsız veri kaynağında performansın korunması ayrı bir değer taşımaktadır. Bu bağlamda, çalışma $R^2 = 0.73$ (bağımsız kaynaklarda) sonucumuz "gerçek dünya/transfer" açısından güçlü bir argüman sunmaktadır.

Ayrıca literatürde sıklıkla vurgulanan bir diğer sorun, ilaç yanıt verilerinin farklı veri setleri arasında değişkenlik gösterebilmesidir. Bu çalışmada model performansının bağımsız veri üzerinde de korunması, geliştirilen yaklaşımın sağlamlığını ve genellenebilirliğini desteklemektedir.

Genomik veriler kullanılarak ilaç yanıtının tahmin edilebileceği, özellikle geniş ölçekli hücre hattı projeleri olan CCLE ve GDSC çalışmaları ile gösterilmiştir. Örneğin, Garnett ve ark. (2012) tarafından yürütülen GDSC çalışması, genomik özellikler ile ilaç duyarlılığı arasında güçlü ilişkiler bulunduğunu ortaya koymuştur. Benzer şekilde, Barretina ve ark. (2012) tarafından geliştirilen CCLE veri tabanı, gen ekspresyonu ve mutasyon profillerinin ilaç yanıtını öngörmede kullanılabileceğini göstermiştir. Bununla birlikte, tek veri kaynağına dayanan çalışmaların biyolojik yorumlama gücü ve genellenebilirliği sınırlı kalabilmektedir.

Daha yakın tarihli çalışmalar, çoklu veri entegrasyonunun tahmin başarımını

artırdığını ve biyolojik yorumlanabilirliği güçlendirdiğini göstermektedir. Örneğin, Iorio ve ark. (2016), farklı veri tabanlarının entegrasyonu ile farmakogenomik modellerin doğruluğunun arttığını bildirmiştir. Ayrıca, son yıllarda açıklanabilir makine öğrenmesi yaklaşımlarının kullanımı artmış ve yalnızca tahmin doğruluğu değil, aynı zamanda biyolojik olarak anlamlı değişkenlerin belirlenmesi de önem kazanmıştır (Lundberg & Lee, 2017).

Bunun yanında, literatürde ilaç yanıt verilerinin farklı veri setleri arasında değişkenlik gösterebildiği ve bu durumun model genellenebilirliğini etkileyebileceği belirtilmektedir (Haibe-Kains ve ark., 2013). Bu çalışmada model performansının bağımsız veri üzerinde korunması, geliştirilen yaklaşımın sağlamlığını ve güvenilirliğini desteklemektedir.

Bu çalışmanın temel güçlü yönlerinden biri çoklu veri entegrasyonu yaklaşımının benimsenmiş olmasıdır. İlaç duyarlılığı verileri, somatik mutasyon profilleri ve gen–ilaç etkileşim bilgileri birlikte kullanılarak tek veri kaynağına dayalı modellere kıyasla daha zengin ve biyolojik açıdan daha gerçekçi bir analiz çerçevesi oluşturulmuştur. Bu bütüncül yaklaşım, modelin hem tahmin gücünü hem de biyolojik anlamlılığını artırmıştır.

Çalışmanın bir diğer önemli güçlü yönü, yüksek model performansının bağımsız veri üzerinden doğrulanmış olmasıdır. Özellikle R^2 metriği ile yapılan değerlendirmeler, modelin açıklayıcılık gücünün yüksek olduğunu ve farklı veri setleri üzerinde performansını koruyabildiğini göstermiştir. Bu durum, geliştirilen yaklaşımın güçlü genellenebilirliğe sahip olduğunu ortaya koymaktadır.

Bunun yanı sıra çalışmada biyolojik yorumlanabilirlik ön planda tutulmuştur. Öznitelik önem analizi ve ablasyon çalışmaları gerçekleştirilmiş, ilaç yanıtında belirleyici rol oynayan kritik genetik değişkenler ve potansiyel belirteçler ortaya konmuştur. Böylece model yalnızca yüksek doğruluklu tahminler üretmekle kalmamış, aynı zamanda biyolojik olarak anlamlandırılabilir sonuçlar sağlamıştır. Çalışma ayrıca sistematik bir analiz yaklaşımı sunmaktadır. Gen–ilaç etkileşim ağları incelenmiş, mutasyon yükü ile etkileşim yoğunluğu birlikte değerlendirilerek kapsamlı bir analiz gerçekleştirilmiştir. Bu yapı, ilaç yanıt mekanizmalarının daha bütüncül biçimde anlaşılmasına katkı sağlamaktadır.

Geliştirilen modelin ilaç yanıtını öngörebilme kapasitesi, kişiselleştirilmiş tedavi

stratejilerinin geliştirilmesine potansiyel katkı sunmaktadır. Bu yönüyle çalışma, hasta bazlı tedavi planlamasına temel oluşturabilecek öngörüsel bir çerçeve ortaya koymaktadır.

Bu çalışmanın bazı sınırlılıkları bulunmaktadır. Öncelikle analizler hücre hattı verilerine dayanmaktadır ve klinik hasta verilerini doğrudan yansıtmamaktadır. Bu nedenle elde edilen bulguların klinik ortama aktarılabilmesi için hasta verileri ile doğrulanması gerekmektedir. Ayrıca açık veri tabanlarından elde edilen verilerde eksiklik, gürültü ve heterojenlik bulunabilmekte olup bu durum model performansını etkileyebilecek potansiyel bir faktör olarak değerlendirilmektedir.

Çalışma ağırlıklı olarak somatik mutasyon verisine dayalıdır ve gen ekspresyonu, metilasyon, kopya sayısı değişimleri (CNV) ve proteomik gibi diğer omik katmanlarının dahil edilmemiş olması biyolojik kapsamı sınırlamaktadır. Bunun yanında, tüm ilaç yanıt mekanizmaları modele yansıtılamamaktadır. Tümör mikroçevresi, epigenetik düzenlemeler ve direnç gelişim mekanizmaları gibi karmaşık biyolojik süreçler modelin dışında kalmıştır.

Geliştirilen model yüksek tahmin doğruluğu ve biyolojik yorumlanabilirlik sunmakla birlikte, klinik uygulamaya doğrudan aktarılabilir nitelikte değildir. Klinik geçerliliğin sağlanabilmesi için modelin hasta verileri üzerinde doğrulanması ve çoklu omik veri entegrasyonu ile desteklenmesi gerekmektedir. Prospektif ve klinik temelli çalışmalar, elde edilen bulguların klinik faydasını ortaya koymak açısından önem taşımaktadır.

Çalışmanın ilk aşamasında küçük hücreli akciğer kanseri (SCLC) veri seti üzerinde mutasyon yüküne dayalı öznitelikler kullanılarak IC50 değeri tahmin edilmiştir. XGBoost, CatBoost ve SVM algoritmalarının karşılaştırıldığı analizlerde en yüksek performansın XGBoost modeli tarafından elde edildiği görülmüştür (RMSE = 1.60; $R^2 \approx 0.60$; MAE = 1.23). Öznitelik önem analizine göre nonsense, frameshift, toplam mutasyon yükü, upstream ve start-loss mutasyonları en belirleyici değişkenler olarak öne çıkmıştır. Bu bulgu, bazı durumlarda mutasyon yüküne ilişkin özelliklerin driver mutasyonlardan daha güçlü öngörü sağlayabileceğini göstermiştir. Model, deneysel çalışmalar öncesinde potansiyel ilaç–hücre etkileşimleri hakkında ön bilgi sunarak karar destek aracı niteliği taşımaktadır.

Çalışmanın ikinci aşamasında ise gen–ilaç–mutasyon verileri entegre edilerek

daha kapsamlı bir farmakogenomik veri seti oluşturulmuş ve çoklu veri entegrasyonuna dayalı bir makine öğrenmesi modeli geliştirilmiştir. Bu model ilk değerlendirmede $R^2 = 0.91$ düzeyinde yüksek açıklayıcılık göstermiş, bağımsız veri ile doğrulamada ise $R^2 = 0.73$ değerini koruyarak güçlü genellenebilirlik ortaya koymuştur. Analizler, ilaç duyarlılığını en fazla etkileyen faktörlerin kritik driver gen mutasyonları, gen–ilaç etkileşim yoğunluğu ve toplam mutasyon yükü olduğunu göstermiştir. Çoklu veri entegrasyonu yaklaşımı, hem tahmin performansını artırmış hem de biyolojik olarak anlamlı farmakogenomik belirteçlerin ortaya konmasına katkı sağlamıştır.

Genel olarak, her iki aşamada da ağaç tabanlı makine öğrenmesi modellerinin ilaç duyarlılığı tahmininde güçlü performans sergilediği görülmüş; özellikle çoklu veri entegrasyonuna dayalı yaklaşımın hem doğruluk hem de biyolojik yorumlanabilirlik açısından daha kapsamlı sonuçlar sunduğu ortaya konmuştur.

Mutasyon yükü bilgilerine dayanarak ilaç duyarlılığının tahmin edilebilmesi için çeşitli makine öğrenmesi yöntemleri uygulanmış ve farklı modeller geliştirilmiştir. CarBoost, XGBoost, SVM ve TabNet algoritmalarıyla oluşturulan modeller, ilaç duyarlılığını %60'ın üzerinde bir doğrulukla tahmin edebilmiştir. Çalışmanın bir diğer amacı ise, mutasyon yükünün belirlenmesinde mutasyon türlerinin etkisini ortaya koymaktır. Bu kapsamda, XGBoost tabanlı makine öğrenmesi modeli ile yapılan analizler; Nonsense, Frameshift, toplam mutasyon sayısı, Upstream, Start Loss ve Driver mutasyonlarının, ilaç duyarlılığı açısından daha belirleyici olduğunu göstermiştir.

Bununla birlikte, hücre hattındaki belirli mutasyon türlerinin niceliksel olarak yüksek sıklıkta bulunması, bu mutasyonların modeldeki ağırlığını artırarak sonuçlar üzerinde etkili olabilmektedir. Bu nedenle, veri setinin genişletilmesi ve farklı hastalık gruplarının eş zamanlı olarak değerlendirilmesi, model performansını daha da artırabilir.

Modelin yüksek doğrulukla tahmin ettiği ilaçların daha ileri düzey deneysel çalışmalar için öncelikli hale getirilmesi, bu çalışmanın potansiyel ilaç adaylarının belirlenmesine katkı sağladığını göstermektedir. Dolayısıyla, mutasyon yükü ve alt bileşenlerinin gelecekte daha kapsamlı bir şekilde ele alınması, çalışmanın önemli bir çıktısı olacaktır.

Günümüzde onkoloji alanında hassas tıp uygulamalarını ilerletmeye yönelik etkili stratejilerden biri, yeni tanımlanan ilaç duyarlılıklarının sistematik biçimde incelenmesidir. Bu bağlamda, kanser hücre hatlarında yapılan ilaç yanıtı taramaları, genetik ve farmakolojik özelliklere dayalı yüksek doğrulukta öngörüler elde etmeye olanak tanımaktadır. Önceki araştırmalar sıklıkla ilaç yanıtını tahmin etmek veya sınıflandırmak amacıyla işlevsel ya da genomik özelliklerden yararlanmıştır.

Bu çalışmada ise, genetik veriler kullanılarak hem ilaçlara hem de hücre hatlarına ait özelliklerin birlikte öğrenilmesini sağlayan çift katmanlı güçlendirilmiş işbirlikçi konu regresyonu (DS-CTR) modeli önerilmektedir. Bu model, ilaç-hücre hattı benzerlik ağı ile grafik modelini birleştiren özgün bir yapıya sahiptir. Kanserde İlaç Duyarlılığının Genomiği (GDSC) projesinden alınan referans veri seti üzerinde gerçekleştirilen 5 katlı çapraz doğrulama analizi, DS-CTR modelinin AUC (Eğri Altındaki Alan) ve ROC (Alıcı Operatör Karakteristiği) ölçütlerine göre dört farklı ileri düzey algoritma grubuna kıyasla anlamlı düzeyde üstün performans sergilediğini ortaya koymuştur.

Literatürde daha önce tanımlanmamış hücre-ilaç etkileşimlerinin ortaya çıkarılmasıyla, DS-CTR modelinin hem doğruluğu hem de yenilik potansiyeli kanıtlanmıştır. Bu yönüyle model, klinik öncesi araştırmalarda yeni anti-kanser tedavi adaylarının keşfini hızlandırma ve maliyetleri düşürme açısından önemli katkılar sunmaktadır.

Çalışmamızda, açık kaynaklı biyomedikal veri tabanları kullanılarak kanser hücre hatlarının genetik profilleri ile ilaç duyarlılığı arasındaki ilişki incelenmiştir. Elde edilen bulgular, belirli genetik faktörlerin ilaç duyarlılığı üzerinde doğrudan etkili olduğunu ortaya koymuştur. Çalışmanın sonuçları, onkoloji ve genetik alanındaki önceki araştırmalarla büyük ölçüde tutarlıdır. Bununla birlikte, elde edilen verilerin klinik uygulamalara nasıl aktarılabileceği, çalışmanın güçlü ve zayıf yönleri ve gelecekteki araştırmalara dair öneriler tartışılacaktır.

Kanser hastalığının tedavisinde ilaç duyarlılığı, hem hasta hem de hücresel düzeyde büyük bir engel oluşturmaktadır. Kanser hücrelerinin, genetik ve epigenetik değişimler yoluyla ilaçlara karşı direnç geliştirebildiği bilinmektedir. Literatürde ilaç duyarlılığı ile ilişkili olan TP53, EGFR, ABCB1 ve KRAS gibi genlerin mutasyonları ve ekspresyon seviyelerinin analiz edildiği çalışmalar mevcuttur. TP53,

hücre siklusunu düzenleyen ve DNA hasarı durumunda apoptozu tetikleyen bir tümör baskılayıcı gen olup, TP53'ün mutasyona uğraması, birçok kemoterapötik ajana karşı direnci artırabilmektedir (Olivier, Hollstein, & Hainaut, 2010). TP53 mutasyonları aynı zamanda potansiyel prognostik ve prediktif belirteçlerin yanı sıra farmakolojik etki hedeflerdir. EGFR geninde meydana gelen belirli mutasyonların, tirozin kinaz inhibitörlerine duyarlılığı artırdığı önceki çalışmalarla gösterilmiştir (Paez ve ark., 2004). ABCB1 geni, hücre zarında bulunan ve ilaçları dışarı atan bir taşıyıcı proteini kodlamaktadır. Tümör heterojenliği, kanser tedavisini zorlaştıran önemli bir faktördür. Kanserler çoğunlukla tek bir mutasyona uğramış hücreden gelişmesine rağmen, zaman içinde genetik, epigenetik ve fenotipik açıdan çeşitlenen farklı hücre toplulukları oluştururlar. Kanser hücrelerinin en önemli özelliklerinden biri, hayatta kalma konusunda son derece başarılı olmalarıdır. Yaşamlarını sürdürebilmek için güçlü ve alternatifli sinyal yollarına sahip olduklarından, sitotoksik ilaçlar gibi dış etkenlere kolayca uyum sağlayabilirler. Üstelik bu süreçler sabit değildir; zamanla ve bulundukları ortama göre değişirler. Örneğin, bir tümörün kenarındaki hücreler, merkezindekilerden farklı olabilir ya da mitozdaki hücreler, interfaz evresinde olanlardan tamamen farklı davranabilir. Üstelik rastgele mutasyonlar ve çevresel baskılar nedeniyle kanser hücreleri sürekli bir değişim ve evrim sürecinden geçer (Gottesman, Lavi, Hall, & Gillet, 2016).

Bu bulgular, genetik profilleme ile kanser hastalarına en uygun ilaç kombinasyonlarının belirlenebileceğini ve kişiselleştirilmiş tedavi yaklaşımlarının önemini vurgulamaktadır. Son yıllarda makine öğrenmesi, genetik verilerdeki karmaşık ilişkileri belirlemek ve ilaç yanıtlarını tahmin etmek için yaygın olarak kullanılmaktadır. Literatürdeki çalışmalarda, Random Forest (Rastgele Orman), Destek Vektör Makineleri (SVM) ve Lojistik Regresyon gibi algoritmalar kullanılarak ilaç yanıtı tahmin modelleri oluşturulmuştur. Kullanılan makine öğrenmesi modelleri, genetik ve ekspresyon verilerini kullanarak belirli ilaçlara duyarlılığı %80'e varan doğruluk oranıyla tahmin edebilmiştir (Ali ve ark., 2021).

Hassas onkolojinin amacı, hastaların tümörlerinin profillerinin genomik bilgisini kullanarak, hastaya özel bireysel tedavi uyarlamaktır. Kanser hücre hatları gibi genetik veri kümeleri, hassas onkolojinin çok önemli bir görevi olan ilaç duyarlılığı tahmini için en değerli kaynaklardır. Büyük kanser hücre hattı panelleri için çoklu

omik verilere dayalı ilaç duyarlılığını tahmin etmek için makine öğrenme yöntemleri kullanılmaktadır. Ancak ilaç duyarlılığı tahmini için makine öğrenimi modellerinin nasıl düzgün bir şekilde eğitileceği ve doğrulanacağına ilişkin kapsamlı yönergeler bulunmamaktadır.

Yapılan bir çalışmada (Xia ve ark., 2022) verilen mevcut halka açık veri kümelerinde ilaç yanıtı tahmini için performans aralığı araştırılmıştır. Beş hücre hattına odaklanarak gözlemlenen yanıt değişkenliğine dayalı çalışma içi ve çalışmalar arası tahmin üst sınırları tarama çalışmaları yapılmıştır. Farklı ilaç yanıtı metriklerini entegre ettikten sonra, modellerin belirli bir konuda nasıl eğitildiğini ölçmek için makine öğrenimi uygulanmıştır. Bu deneyler veri kümelerinde daha öngörülebilir olan veya tahmin edilebilir değerleri olanları içerir. Yeni deneysel verilerin değerini anlamak ve göreceli olarak incelemek için, ilaç çeşitliliğine karşı hücre hattının önemi, modelin genellenebilirliğini etkileyebilen simülasyonlar kullanılmıştır. Çalışmada kullanılan beş farklı ilaç-yanıt veri setleri NCI (The National Cancer Instute) ve PharmacoDB' den indirilmiştir. Bu beş farklı veri setlerinin önemli ölçüde örtüştüğü görülmüştür. Çalışılacak örnek koleksiyonu seçimi oluşturmak için önce hücre hattı seti ve ilaç seti tanımlanmıştır. Diğer veri setlerdeki tüm ilaçlar çalışmaya dahil edilerek, toplamda 1801 ilaç seçilmiştir. Hücre dizisi tabanlı antikanser taramasında, ilaç duyarlılığı belirli bir konsantrasyonda ilaç tedavisine maruz kaldıktan sonra hücrenin hayatta kalma oranını ölçen doz-yanıt değerleri ile ölçülür. Doz-yanıt değerleri, belirli bir hücre-ilaç çifti için ilaç yanıtının dozdan bağımsız ölçümleri için özetlenebilir. Beş çalışmanın tümü farklı aralıklar ile doz-yanıt verileri sağlamıştır. Dozdan bağımsız metrikler daha karmaşık izlenmiştir. Örneğin, GDSC IC50 kullanmışken (%50 hücre yaşayabilirlik inhibisyonu dozu), NCI60, GI50 kullanmıştır (maksimum değerin %50'sinin büyüme inhibisyonu gözlenir).

Bilimsel uzlaşı olmadığından, toplanmış verilerin yeniden hesaplanması için her bir doz-yanıt değerlerindeki bağımsız metriklerin hesaplamadaki tutarsızlıklarının giderilmesi ve eğri uyumlama yapılarak PharmacoDB' nin uygulaması benimsemiştir. Bu istatistikler arasında AAC (area above the dose–response curve) test edilen ilaç konsantrasyonları için doz-yanıt eğrisi ve DSS (ilaç duyarlılık skoru) doz aralığına karşı daha sağlam metriklerdir. Farklı çalışmalardan elde edilen veri

kümeleri genellikle bilimsel yanlılık sergileyebilirler. Çalışma tasarımındaki farklılıklar dışında (örn. hücre dizileri ve ilaç seçimi, örneklem büyüklüğü) deneysel olarak ölçülen veriler ile ilişkili moleküler karakterizasyonda ve yanıt analizinde iki ana yanlılık türü bulunmaktadır.

Yapılan çalışmada makine öğrenimi sorunu olarak girdi karakterlerinde ve çıktı etiketlerinde zorluklar gözlenmiş olup birleştirmenin ana zorluğunu oluşturmaktadır. Öğrenme ve öğrenme çabalarını çoklu veri kaynakları üzerinden aktarma da bir diğer zorluk olarak görülmüştür. İlk yanlılık türü, özellik, ön işleme ile hafifletilebilmektedir. Spesifik olarak, farklı veritabanlarından gen ifade profilleri ve izlenen ilaç özelliklerini homojenleştirmek için çalışmada toplu düzeltme (batch correction) protokolü kullanılmıştır. İkinci yanlılık türü bu makalenin odak noktasını oluşturmuştur. Entegrasyon çabalarına rağmen ölçülen ilaçta yanıt farklılıkları arasında önemli uyumsuzluklar kalmıştır. Çalışmalar arası yanıt değişkenliğini ve farklı çalışmalarda ilaç yanıtını karşılaştırmak için, her çalışma genellikle kendi deneysel konsantrasyonlarına sahip olduğundan, doza bağımlı metrikler kullanılmaktadır. Bunun yerine çalışmada üç dozdan bağımsız yanıt metriği: AUC, AAC ve DSS tercih edilmiştir. AUC ve AAC arasındaki temel fark, AUC' nin sabit bir doz aralığına dayalı olduğu, AAC' nin ise belirli bir ekranda görüntülenen ilaca özgü doz penceresine dayandığıdır. DSS, her iki ilaç etkinliğini birleştiren başka bir ölçümdür.

Bu üç ölçümü kullanarak, çapraz çalışma yanıtı değişkenlikleri çoklu çalışmalarda görünen hücre hattı- ilaç çiftlerinde analiz edilmiştir. Çalışmada doz yanıtını tahmin etmek için makine öğrenimi modelleri uygulanmıştır. Tek bir veri seti bağlamında değerlendirildiğinde, tahmin performansı girdi özellikleri ve model karmaşıklığı genel faktörlerine bağlanmıştır. İncelenen modellerde bu iki faktör üzerinde geniş kapsamlı konfigürasyonlar sunulmuştur. Ancak bu modellerin çoğu, tek bir çalışma içinde çapraz doğrulama ile değerlendirmede durdurulmuştur. Kaynak ve hedef veri kümeleri arasında çapraz çalışma tahmin performansının değerlendirilmesi için artan karmaşıklığa sahip üç makine öğrenimi modeli uygulanmıştır. İlk ikisi Random Forest ve LightGBM olarak kullanılan temel modeller olup, üçüncüsü de bu çalışmayı yapanların oluşturduğu deep neural network metoduydu. Makalede, beş kanser hücre dizisinde çalışmalar içi ve

çalışmalar arasındaki ilaca yanıt değişkenlikleri üzerine odaklanılmıştır. Sistematik analiz yapılarak, modellerin tek bir platform üzerinde nasıl eğitildiği karşılaştırılmıştır. Derin öğrenme sonuçları umut verici olsa da, ilaç yanıtı tahmininin gelecekteki başarısı modelin iyileştirilmesine ve genellenebilirliğine dayanacaktır. İlaç çeşitliliğinde deneysel standardizasyon ve önceliklendirme de verilerin taranması, bütünleştirici öğrenme için veri taramanın değerini en üst düzeye çıkarmak için anahtar olacaktır (Xia ve ark., 2022).

Ahmadi ve Eslahchi (2021) tarafından yapılan bir çalışmada bir sınıflandırma makinesi önerilmiştir. Bir hücre dizisi için duyarlı/dirençli ilaçları öngören öğrenme yaklaşımı hem ilaç hem de hücre hattı benzerliklerini kullanarak, hücre hattı veya ilaç benzerliklerinden birini kullanarak veya herhangi bir benzerlik bilgisi kullanılmasa bile uygulanabilir. Bu çalışmada önerilen yöntem ilacı belirlemek için maksimum konsantrasyon eşiklerini kullanır ve sınıf etiketlerine yönlendirir. Performansı, iki veri setinde hücre hattı-ilaç çiftleri üzerinde katmanlı beş katlı çapraz doğrulama kullanılarak değerlendirilmiştir. Önerilen modelin çalışmada tahmin gücünün değerlendirilmesi ve son teknoloji sınıflandırma yöntemleri de dahil olmak üzere üç yöntem seti, son teknoloji regresyon yöntemleri ve kullanıma hazır sınıflandırma makine öğrenimi yaklaşımlar, önerilen yöntemin diğer yöntemlerden daha iyi performans gösterdiğini tespit etmişlerdir. Ayrıca, modelin etkinliği dokuya özgü koşullarda değerlendirilmiştir ve literatürdeki güvenilir veri tabanları da bu model tarafından tahmin edilen ilişkilendirmeler ile uyumlu bulunmuştur. Bu nedenle, önerilen modelin kanser önleyici ilaç duyarlılığını tahmin etmek için verimli bir şekilde kullanılabileceği savunulmaktadır.

Kanser tedavisinde ilaç duyarlılığını doğru bir şekilde öngörebilmek amacıyla çalışmamızda, bir tahmin yöntemi geliştirilmiştir. Kanser hücrelerinin canlılığını inhibe etme gücünü ölçen yarı maksimal inhibitör konsantrasyonu (IC50) değeri, ilaç etkinliğinin yaygın olarak kullanılan bir göstergesidir. Daha düşük IC50 değerleri, ilacın daha yüksek bir biyolojik etkiye sahip olduğunu göstermektedir.

Hücrelerin tümör mutasyon yüküne (Tumor Mutational Burden, TMB) dayalı olarak çalışma kapsamında, IC50 değerleri tahmin edilmiştir. Tümör mutasyon yükü, kanser hücrelerinin genomunda biriken genetik mutasyonların sayısını ifade eder ve tümör immünojenisitesi ile terapötik yanıtta potansiyel bir belirteç olarak

değerlendirilmektedir. Bu doğrultuda, ilaç yanıtını tahmin etme problemini çözmek için makine öğrenmesi (ML) algoritmaları ve ağ biyolojisi yaklaşımları birlikte kullanılmıştır. İlaç yanıtının önceden öngörülebilmesi, kişiselleştirilmiş tıbbın etkin şekilde uygulanması açısından kritik öneme sahiptir. Geleneksel olarak, araştırmacılar farklı hasta gruplarında ilaç yanıtlarını değerlendirmek için klinik çalışma verilerinden elde edilen tedavi sonuçlarına başvurmaktadır. Ancak bu yaklaşım, etik sınırlamalar ve yüksek maliyetler nedeniyle çeşitli zorluklar içermektedir.

Son yıllarda birçok çalışma, kanser hücre hatlarının genomik profillerini giriş verisi olarak kullanarak genetik bilgileri analiz etmekte ve bu bilgilerden ilaç yanıtı tahmin modelleri geliştirmektedir. Elde edilen bu modeller, hasta verilerinin genomik profilleri ile ilaçla ilişkili hücre hatlarının karşılaştırılması yoluyla, bireylerin farklı ilaçlara vereceği olası yanıtları öngörmek için kullanılmaktadır.

Genetik veri tabanları, ilaç yanıtı tahmin modellerinin eğitimi için önemli bir kaynak olarak kullanılmaktadır (Barretina ve ark., 2012; Yang ve ark., 2013). Bu modelleri sınıflandırmak amacıyla genellikle iki temel yaklaşım benimsenmektedir. Bunlar tekli veya çoklu omik tabanlı modeller ve ilacın kimyasal özellikleri ile desteklenen modellerdir (Park ve ark., 2022).

Tekli veya çoklu omik tabanlı yöntemler, genellikle belirli bir ilaca yönelik olarak, kanser hücre hatları ile birlikte bir ya da daha fazla genomik veri türünü (örneğin gen ekspresyonu, DNA metilasyonu, mutasyon profilleri ve kopya sayısı değişimleri) kullanarak özelleştirilmiş tahmin modelleri oluşturmayı amaçlamaktadır (Bomane, Gonçalves, & Ballester, 2019; Chiu ve ark., 2019; Costello ve ark., 2014; Wei, Liu, Zheng, & Li, 2019).

Önceki araştırmalarda, derin öğrenme modellerinin doğruluğunun, modelin sahip olduğu katman sayısı ve yapısal derinliği ile paralel olarak arttığı belirtilmiştir (Lecun, Bottou, Bengio, & Haffner, 1998; Vestias, 2020). Nitekim bir çalışmada, farklı omik veri türleri ve değişen veri seti boyutları altında IC50 değerlerini tahmin etmede regresyon tabanlı makine öğrenmesi (ML) ve derin öğrenme (DL) modelleri kapsamlı biçimde karşılaştırılmış; çoğu durumda DL modellerinin ML modellerine göre daha üstün performans sergilediği ortaya konmuştur (Park ve ark., 2022).

İntratümöral malignitelerin yüksek genetik heterojeniteye sahip olması, ilaç

yanıtlarında bireysel farklılıklara yol açmaktadır. Bu durum, kişiselleştirilmiş tıbbın başarısı için ilaç yanıtı tahmininin ne denli önemli olduğunu ortaya koymaktadır. Ancak, şimdiye kadar geliştirilen bazı modellerde ne mutasyon durumu ne de ilaçların moleküler "parmak izi" özellikleri yeterince dikkate alınmamıştır. Bu eksikliği gidermek amacıyla geliştirilen DeepIC50 adlı model, 27.756 özelliği kapsayacak şekilde ilaçlara ait kimyasal parmak izleri ve hücrelerin mutasyon durumlarını kullanarak üç farklı ilaç yanıt sınıfını tahmin etmeyi hedeflemiştir. 1 boyutlu konvolüsyonel sinir ağı (1D-CNN) mimarisiyle geliştirilen bu model, hücre canlılığı ve IC50 tahmini açısından iki farklı pan-kanser hücre hattı veri setinde üstün performans göstermiştir. DeepIC50, yeni bileşiklerin çeşitli kanser hücre hatları üzerindeki etkilerini öngörme potansiyeli taşıyarak ilaç keşif süreçlerinde önemli katkılar sunmaktadır (Joo ve ark., 2019).

Büyük ölçekli kanser hasta kohortlarında doğrudan ve kesin tedavi yanıtı verilerinin elde edilmesi son derece güçtür. Bu nedenle, genellikle hücre hatları ve fare modelleri gibi preklinik sistemler, kanser farmakogenomiği alanındaki keşiflerin temel kaynakları olarak kullanılmaktadır. Ancak bu deneysel modeller, özellikle sınırlı örneklem büyüklükleri gibi çeşitli kısıtlamalara sahiptir.

Öte yandan, Kanser Genom Atlası (The Cancer Genome Atlas, TCGA) gibi geniş kapsamlı klinik genomik veri setleri, hasta bazlı analizlerde daha zengin bir bağlam sunmaktadır. Bu tür veri setlerinde, spesifik ilaç yanıtlarının doğru bir şekilde belirlenmesi kritik önem taşımaktadır. Bu bağlamda geliştirilen yaklaşım, öncelikle gen ekspresyon profillerine dayanarak geniş bir kanser hücre hattı panelinde tedavi yanıtını modelleyen istatistiksel yöntemlerin oluşturulmasını içermektedir. Ardından, bu modeller, klinik veri setlerinde yer alan tümörlere ait gen ifade profilleri üzerinde uygulanarak bireysel hastalar için öngörülen farmakolojik yanıtlar üretmektedir.

Elde edilen bu tahmini ilaç yanıtı verileri, ilgili tümörlerde gözlemlenen somatik genetik varyasyonlarla ilişkilendirilerek analiz edilmektedir. Bu tür modelleme yaklaşımları, kişiselleştirilmiş tedavi stratejilerinin geliştirilmesinde önemli bir araç olarak öne çıkmakta ve gen ekspresyonu ile ilaç duyarlılığı arasındaki ilişkilerin klinik düzeyde yorumlanmasını mümkün kılmaktadır (Geeleher ve ark., 2017).

Bu çalışma, kanser tedavisinde kişiselleştirilmiş tıp uygulamalarını destekleyen

genetik modellemelere katkı sunmayı amaçlamaktadır. Açık erişimli büyük ölçekli veri setlerinin kullanılması, çalışmanın şeffaflığını ve tekrarlanabilirliğini artırarak bilimsel güvenilirliğini güçlendirmiştir. Makine öğrenmesi algoritmaları ile istatistiksel analizlerin birlikte uygulanması, ilaç yanıtı tahminlerinin doğruluğunu önemli ölçüde artırmıştır. Ayrıca, farklı kanser türlerine ait hücre hatlarının analiz edilmesi, geliştirilen modellerin farklı klinik senaryolarda uygulanabilirliğini göstermekte ve bulguların genellenebilirliğini desteklemektedir.

Bununla birlikte, çalışmanın bazı sınırlılıkları da bulunmaktadır. İn vitro hücre hatlarına dayalı analizlerin klinik geçerliliği sınırlı olabilir; zira laboratuvar ortamında elde edilen veriler, insan organizmasındaki biyolojik karmaşıklığı tam olarak yansıtmayabilir (Barretina ve ark., 2012). Ayrıca, çalışmada yalnızca genetik ve gen ekspresyonu verilerine odaklanılmış olup, epigenetik faktörler, tümör mikroçevresi etkileri ve bağışıklık sistemi gibi önemli biyolojik değişkenler değerlendirme dışında bırakılmıştır.

Elde edilen bulgular, ilaç duyarlılığı ile genetik profiller arasındaki ilişkinin anlamlandırılmasında makine öğrenmesi temelli modellerin etkinliğini ortaya koymaktadır. Özellikle mutasyon yükü ile ilaç yanıtı arasındaki ilişkinin saptanması, klinisyenlerin belirli genetik profillere sahip hastalar için en uygun ilaç seçiminde karar destek aracı olarak bu tür modellerden yararlanabileceğini göstermektedir. Ancak bu modellerin klinik uygulamalarda güvenle kullanılabilmesi için bazı temel gereklilikler bulunmaktadır. Bunlar arasında, geliştirilen tahmin modellerinin klinik düzeyde doğrulanması, hücre hatları yerine doğrudan hastadan alınan primer örneklerin veya üç boyutlu organoid modellerin kullanılması, epigenetik ve immünogenomik etkileşimlerin analize dâhil edilmesi ve daha büyük, çeşitli hasta veri setleriyle modellerin yeniden eğitilmesi sayılabilir. Ayrıca, farmakogenetik testlerin klinik çalışmalarda yaygınlaştırılması, bireylerin genetik profiline dayalı kişiselleştirilmiş tedavi stratejilerinin rutin tıpta uygulanabilirliğini artıracaktır.

Bu çalışma, açık kaynaklı veriler ve makine öğrenmesi yaklaşımları kullanılarak kanser hücre hatlarında ilaç yanıtının genetik temellerinin analiz edilebileceğini göstermiştir. Biyoinformatik ve yapay zekâ temelli modellemelerin, genetik verilerin yorumlanmasında güçlü araçlar sunduğu ortaya konulmuştur. Bununla birlikte, klinik geçerliliğin artırılması ve model sonuçlarının hasta bazında uygulanabilir hale

getirilmesi için geniş ölçekli, çok merkezli klinik doğrulama çalışmalarına ihtiyaç duyulmaktadır. Bu doğrultuda geliştirilecek yeni nesil genetik modeller, kişiselleştirilmiş tedavi yaklaşımlarının onkoloji pratiğinde yaygınlaşmasına katkı sağlayacaktır.

İlaç etkileşimlerini öngörecek güçlü bir makine öğrenmesi modeli tasarlayabilmek için geniş kapsamlı ve zengin bir veri setine ihtiyaç duyulması önem arzetmektedir. Her ne kadar bu alandaki açık erişimli veri tabanları giderek gelişiyor olsa da, farklı veri kaynaklarının bir araya getirilmesi veri setini daha anlamlı hâle getirmekte ve model eğitimini kolaylaştırmaktadır. Bu çalışmada bu amaçla, GDSC, COSMIC ve DGIdb veri tabanlarından dört farklı veri seti toplanmış ve birleştirilmiştir. GDSC kanser hücre hatlarında kullanılan ilaçlara ait duyarlılık verilerini, COSMIC kanserli dokulardaki somatik mutasyon bilgilerini, DGIdb ise kapsamlı gen–ilaç etkileşimlerini sunmaktadır. Böyle bir çoklu veri entegrasyonu, henüz keşfedilmemiş olabilecek yeni etkileşim desenlerinin yakalanmasına ve bunların ilaç etkileşimlerinde rolünün anlaşılmasına katkı sağlayabilir.

Ancak veri entegrasyonu beraberinde bazı güçlükler de getirmektedir. Veri setinin boyutu hızla büyürken, veri seyrekleşebilmekte ve birçok sütun eksik değerler içerebilmektedir.

Bu sorunları aşmak için veriye ön işleme uygulanması zorunludur. Eksik değerler bilimsel yöntemlerle tamamlanmalı ya da uygun şekilde imputasyon yapılmalıdır. Ayrıca, büyük hacimli entegre veriler, yüksek hesaplama maliyetine veya aşırı öğrenmeye yol açabileceğinden, korelasyon temelli filtreleme yöntemleri kullanılması gerekebilir. Son aşamada, etkili olmayan sütunların ayıklanması için özellik önemliliğine dayalı seçim yöntemleri devreye sokulmalıdır.

Filtreleme adımlarıyla özellik sayısını 27'ye indirgedikten sonra, en bilgilendirici özellikleri seçmek amacıyla optimizasyon algoritmaları uygulanmıştır. Bu aşamanın temel hedefi, 27 özellik arasından makine öğrenmesi için en uygun alt kümenin belirlenmesidir. PSO tabanlı deneyler, biyomedikal verilerin doğru kodlama yöntemleriyle sayısallaştırılması ve özellik seçiminin performansı önemli ölçüde yükselttiğini ortaya koymuştur. Bu yönüyle çalışmamız, literatürde genellikle ayrı ayrı değerlendirilen GDSC, COSMIC ve DGIdb kaynaklarını birlikte kullanan nadir örneklerden biridir.

Önerilen işlem hattı, veri toplama, birleştirme, filtreleme ve model geliştirme aşamalarını sistematik şekilde bir araya getirerek literatüre yenilikçi bir katkı sağlamaktadır. Daha önce birçok çalışmada GDSC, COSMIC ve DGIdb kullanılmış olsa da, bu üç kaynağın tek bir birleşik veri setinde entegre edilmesi ve ardından özellik seçimi ile optimizasyon adımlarının eklenmesi özgün bir yöntemdir. Elde edilen zenginleştirilmiş veri seti, makine öğrenmesi modellerine girdi olarak sunulmuş ve ilaç duyarlılığının etkili şekilde öngörülmesine imkân tanımıştır.

Bununla birlikte, geliştirilen modellerin gerçek klinik ortamda geçerliliğini görmek için ileriye dönük klinik çalışmalarla test edilmesi gerekmektedir. Özellikle farklı hasta gruplarında ve farklı kanser tiplerinde yapılacak validasyon, modellerin genelleme kabiliyetini artıracaktır. Translasyonel tıp açısından, laboratuvar kaynaklı kanser hücre hattı verilerinin klinikte kullanılabilecek öngörü modellerine dönüştürülmesi, araştırma ile hasta bakımını buluşturan kritik bir aşamayı temsil etmektedir. Çalışmamız, bu bağlamda alana yeni bir perspektif kazandırmakta ve translasyonel tıp için önemli bir katkı sunmaktadır.

Genetik araştırmalarda, ilaç yanıtının genetik temellerini kavramak her geçen gün daha kritik hale gelmektedir. Bu bağlamda, geliştirilen modelimiz bu bilgiyi klinik uygulamalara aktarmada bir köprü görevi görebilir. Özellikle belirli mutasyon profillerine sahip kanser türlerinde hangi ilaçların daha etkili olabileceğini öngörmek, ilaç geliştirme süreçlerini hızlandırırken maliyetleri de düşürebilir. Translasyonel tıp açısından bakıldığında, laboratuvar ortamında üretilen kanser hücre hattı verilerinin klinikte kullanılabilir öngörü modellerine dönüştürülmesi, araştırma ile hasta bakımı arasındaki boşluğu kapatacak kritik bir adım olarak değerlendirilebilir.

Bununla birlikte, çalışmanın bazı sınırlılıkları vardır. Mevcut açık erişimli ilaç duyarlılığı veri tabanları, yüksek doğrulukta tahminler yapabilmek için hâlâ sınırlı kalmaktadır. İlaç–kanser duyarlılığı ilişkisini ortaya koyma konusu son dönemde önemli bir ilgi alanı olsa da, farklı veri tabanlarının sunduğu içerik çeşitliliği entegrasyonu zorlaştırmaktadır. Örneğin, GDSC kanser hücre hatlarının ilaç yanıtını; COSMIC kanser dokularındaki somatik mutasyonları; DGIdb ise gen–ilaç etkileşimlerini sunmaktadır. Ayrıca, yeni ilaçlardan elde edilen deneysel bulgular hesaplama süreçlerini daha karmaşık hale getirebilmektedir. Dolayısıyla, geliştirilen

makine öğrenmesi modellerinin aşırı öğrenmeye (overfitting) karşı dayanıklı olması ve güçlü genelleme kapasitesi sunması için daha fazla geliştirme gerekmektedir.

Bu sınırlılıkların aşılması ve bulguların güçlendirilmesi için, gelecekte daha geniş ve daha çeşitli veri setlerinin kullanılması gereklidir. Böylece modellerin klinikte daha farklı hasta gruplarında uygulanabilirliği artacaktır. Ayrıca, genomik verilerin yanında transkriptomik, proteomik, metabolomik ve epigenomik verilerin de entegre edilmesi, ilaç yanıtını şekillendiren faktörlerin daha kapsamlı biçimde anlaşılmasına katkı sağlayacaktır.

Ayrıca, gen, mutasyon ve ilaç etkileşimi veri tabanları sürekli olarak güncellenmekte ve yeni araştırma grupları tarafından yeni veri kaynakları ortaya konmaktadır. Açık erişimli veri tabanları büyüyüp çeşitlendikçe, geliştirdiğimiz çerçevenin de bu yeni verilerle güncellenmesi gerekecektir. Gelecekte, yeni veri tabanlarının entegrasyonu ve multi-omik verilerin dahil edilmesi düşünülebilir. Bunun yanında, bu çalışmanın devamı olarak etkileşimli araçlar ve kullanıcı dostu arayüzlerin geliştirilmesi de önem kazanmaktadır.

Bu çalışma deneysel bulgular ile hesaplamalı genetik analizler arasında bir köprü oluşturarak, translasyonel tıp alanına önemli bir katkı sunmaktadır.

Tez çalışmamın ikinci kısmında, açık kaynaklı ilaç–gen–yolak–mutasyon veri tabanları üzerinde kategorileştirme ve filtreleme adımları uygulanarak yeni bir ilaç duyarlılığı/direnç veri seti oluşturulmuştur. Veri zenginleştirme sonrası yapılan istatistiksel analizler, ilaç duyarlılığı ile ilişkili en yoğun etkileşimde bulunan gen ve ilaçları ortaya çıkarmıştır. Bu veriler ışığında, makine öğrenmesi algoritmaları geliştirilmiş ve ilaç duyarlılığının öngörülmesinde kullanılabilir hale getirilmiştir. Çalışma, gen–mutasyon–ilaç etkileşimleri ile ilaç yanıtı arasında güçlü bağların kurulabileceğini göstermiştir. CCLE'den elde edilen kapsamlı kopya sayısı değişimi (CNV) ve mutasyon profilleri, hem genomik toplu özellikler hem de dağılımsal kodlamalar kullanılarak temsil edilmiştir. Ayrıca, ilaç ve doku bilgileri, bilgi sızıntısını önlemek amacıyla yalnızca eğitim verisi üzerinden öğrenilen kontrollü kodlama stratejileri aracılığıyla modele dahil edilmiştir.

Daha sonra, geniş bir şekilde oluşturulmuş özellik havuzundan Parçacık Sürü Optimizasyonu (PSO) yöntemi kullanılarak bilgilendirici sayısal özelliklerden oluşan bir alt küme seçilmiş ve nihai tahmin modeli LightGBM algoritması kullanılarak

eğitilmiştir. CTRP–CCLE veri entegrasyonu üzerinde elde edilen performans, önerilen analiz hattının farklı özellik yapıları ve farklı ilaç duyarlılığı tanımlarına sahip alternatif farmakogenomik veri setlerine de etkili bir şekilde uyarlanabileceğini göstermektedir.

Kanser mutasyonları, hücre hatları, yolaklar ve ilaç etkileşimlerini tek bir çatı altında birleştirmek için etkili bir makine öğrenmesi pipeline tasarımı kritik öneme sahiptir. Farklı veri tabanlarının birleştirilmesi sonrası ortaya çıkan büyük veri hacmi, doğru filtreleme ve özellik seçimi adımlarıyla düzenlenmelidir. Bu bağlamda özellik kardinalitesi temel bir ölçüt olarak öne çıkmaktadır. Ayrıca, optimizasyon algoritmaları en bilgilendirici özellikleri belirlemede önemli katkı sağlamaktadır.

İlaç duyarlılığını öngörmede makine öğrenmesi vazgeçilmez bir araçtır ve her algoritma için hiperparametre optimizasyonu kaçınılmazdır. Çalışmada geliştirilen modeller, aşırı öğrenmeye düşmeden yüksek doğrulukta tahminler sunabilmiştir. Gen mutasyonları ile ilaç duyarlılığı arasındaki ilişkinin anlaşılması, kişiselleştirilmiş tedavi stratejilerinin geliştirilmesinde temel bir adımdır. Günümüzde veri kaynaklarının sınırlılığı nedeniyle yapay zekânın katkıları kısıtlı olsa da, veri miktarının artmasıyla birlikte kişiselleştirilmiş tıp alanında daha güçlü ve etkin sonuçlar alınabilecektir.

Geliştirdiğimiz model ilaç duyarlılığını yüksek doğrulukla tahmin etmiş, genetik mutasyonların ilaç yanıtındaki belirleyici rolünü ortaya koymuş ve kişiselleştirilmiş tedaviye katkı sağlayabilecek güçlü bir farmakogenomik çerçeve sunmuştur.

Gelecekte yapılacak çalışmalar kapsamında modelin klinik hasta verileri üzerinde doğrulanması ve prospektif validasyon çalışmalarının gerçekleştirilmesi hedeflenmektedir. Ayrıca, çoklu omik veri entegrasyonu yaklaşımı genişletilerek RNA-seq, kopya sayısı değişimleri (CNV), DNA metilasyonu ve proteomik verilerin modele dahil edilmesi, biyolojik kapsamın artırılmasına katkı sağlayacaktır. Daha güçlü ve karmaşık örüntüleri yakalayabilen derin öğrenme, grafik sinir ağları ve topluluk (ensemble) temelli gelişmiş yapay zekâ yaklaşımlarının kullanılması, tahmin performansının daha da iyileştirilmesine olanak sağlayabilir.

Bunun yanı sıra yeni yapay zekâ yöntemlerinin uygulanması, kritik biyobelirteçlerin belirlenmesine ve model çıktılarının klinik açıdan

yorumlanabilirliğinin artırılmasına katkı sunacaktır. Gelecekte ilaç direnci ve kombinasyon tedavisi yanıtını öngörebilen modellerin geliştirilmesi de önemli bir araştırma alanı olarak öne çıkmaktadır. Son olarak, geliştirilen yaklaşımın klinik kullanıma aktarılabilmesi amacıyla, hekimlere karar destek sağlayacak kullanıcı dostu bir yazılım veya arayüz geliştirilmesi planlanmaktadır.

6. KAYNAKLAR

Abbas, S., Sampedro, G. A., Abisado, M., Almadhor, A. S., Kim, T.-H., & Zaidi, M. M. (2023). A novel drug-drug indicator dataset and ensemble stacking model for detection and classification of drug-drug interaction indicators. IEEE Access, 11, 101525–101536. https://doi.org/10.1109/ACCESS.2023.3315241

Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B., & Goldenberg, A. (2020). Machine learning approaches to drug response prediction: Challenges and recent progress. NPJ Precision Oncology, 4, 19. https://doi.org/10.1038 /s41698-020-0122-1

Ahmadi Moughari, F., & Eslahchi, C. (2021). A computational method for drug sensitivity prediction of cancer cell lines based on various molecular information. PLOS ONE, 16(4), e0250620. https://doi.org/10.1371/journal.pone.0250620

Ali, M. M., Ahmed, K., Bui, F. M., Paul, B. K., Ibrahim, S. M., Quinn, J. M. W., & Moni, M. A. (2021). Machine learning-based statistical analysis for early stage detection of cervical cancer. Computers in Biology and Medicine, 139, 104985. https://doi.org/10.1016/j.compbiomed.2021.104985

Alizadehsani, R., Oyelere, S. S., Hussain, S., Jagatheesaperumal, S. K., Calixto, R. R., Rahouti, M., ... & De Albuquerque, V. H. C. (2024). Explainable artificial intelligence for drug discovery and development: a comprehensive survey. IEEE Access, 12, 35796-35812.

Alpaydin, E. (2020). Introduction to machine learning (4th ed.). Cambridge, MIT Press.

Asfari, M., Janjic, D., Meda, P., Li, G., Halban, P. A., & Wollheim, C. B. (1992). Establishment of 2-mercaptoethanol-dependent differentiated insulin-secreting cell lines. Endocrinology, 130(1), 167–178. https://doi.org/10.1210/endo.130.1.1370150

Ashburn, T. T., & Thor, K. B. (2004). Drug repositioning: Identifying and developing new uses for existing drugs. Nature Reviews Drug Discovery, 3(8), 673–683. https://doi.org/10.1038/nrd1468

Baik, C. S., Rubin, E. H., Forde, P. M., Mehnert, J. M., Collyar, D., Butler, M. O., …. Chow, L. Q. M. (2017). Immuno-oncology clinical trial design: Limitations, challenges, and opportunities. Clinical Cancer Research, 23(17), 4992–5002. https://doi.org/10.1158/1078-0432.CCR-16-3066

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., … Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 483(7391), 603–607. https://doi.org/10.1038/nature11003

Batlle, E., & Clevers, H. (2017). Cancer stem cells revisited. Nature medicine, 23(10), 1124–1134. https://doi.org/10.1038/nm.4409

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In Noise reduction in speech processing (pp. 1–4). Springer. https://doi.org/10.1007/978-3-642-00296-0_5

Bergstra, J. and Bengio, Y. (2012) Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research, 13, 281-305.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J.,

… Meyerson, M. (2010). The landscape of somatic copy-number alteration across human cancers. Nature, 463(7283), 899-905. https://doi.org/10.1038/nature08822

Bomane, A , Gonçalves, A., Ballester, P.J. (2019). Paclitaxel Response Can Be Predicted With Interpretable Multi-Variate Classifiers Exploiting DNA-Methylation and miRNA Data. Front Genet. 10:1041. https://doi.org/10.3389/fgene.2019.01041

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 74(3), 229–263. https://doi.org/10.3322/caac.21834

Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Brown, J. M., Campbell, J. P., Beers, A., Chang, K., Ostmo, S., Chan, R. V. P., … Imaging and Informatics in Retinopathy of Prematurity (i-ROP) Research Consortium (2018). Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. JAMA ophthalmology, 136(7), 803–810. https://doi.org/10.1001/jamaophthalmol.2018.1934

Bueschbell, B., Caniceiro, A. B., Suzano, P. M. S., Machuqueiro, M., Rosário-Ferreira, N., & Moreira, I. S. (2022). Network biology and artificial intelligence drive the understanding of the multidrug resistance phenotype in cancer. Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy, 60, 100811. https://doi.org/10.1016/j.drup.2022.100811

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., … Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics, 45(10), 1113–1120. https://doi.org/10.1038/ng.2764

Carli, F., Di Chiaro, P., Morelli, M., Arora, C., Bisceglia, L., De Oliveira Rosa, N.,… Raimondi, F. (2025). Learning and actioning general principles of cancer cell drug sensitivity. Nature Communications, 16(1), 1654. https://doi.org/10.1038/s41467-025-56827-5

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16), 785–794. https://doi.org/10.1145/2939672.2939785

Chiu, Y. C., Chen, H. H., Zhang, T., Zhang, S., Gorthi, A., Wang, L. J., … Chen, Y. (2019). Predicting drug response of tumors from integrated genomic profiles by deep neural networks. BMC Medical Genomics, 12(Suppl 1), 18. https://doi.org/10.1186/s12920-018-0460-9

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J. P. (2020). Introduction to machine learning, neural networks, and deep learning. Translational Vision Science & Technology, 9(2), 14. https://doi.org/10.1167/tvst.9.2.14

Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., & Pritchard, J. K. (2010). A world-wide survey of structural variation. Nature, 464(7289), 713–720. https://doi.org/10.1038/nature08939

Convit, J., Browne, S. G., Languillon, J., Pettit, J. H., Ramanujam, K., Sagher, F., … Martínez Domínguez, V. (1970). Therapy of leprosy. Bulletin of the World

Health Organization, 42(5), 667–672.
Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., … Stolovitzky, G. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. Nature Biotechnology, 32(12), 1202–1212. https://doi.org/10.1038/nbt.2877
Coyner, A. S., Swan, R., Campbell, J. P., Ostmo, S., Brown, J. M., Kalpathy-Cramer, J., … Imaging and Informatics in Retinopathy of Prematurity Research Consortium. (2019). Automated fundus image quality assessment in retinopathy of prematurity using deep convolutional neural networks. Ophthalmology Retina, 3(5), 444–450. https://doi.org/10.1016/j.oret.2019.01.015
De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., … Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine, 24(9), 1342–1350. https://doi.org/10.1038/s41591-018-0107-6
Douglas, E. J., Fiegler, H., Rowan, A., Halford, S., Bicknell, D. C., Bodmer, W., … Carter, N. P. (2004). Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. Cancer research, 64(14), 4817–4825. https://doi.org/10.1158/0008-5472.CAN-04-0328
Drexler, H. G., Matsuo, A. Y., & MacLeod, R. A. (2000). Continuous hematopoietic cell lines as model systems for leukemia-lymphoma research. Leukemia research, 24(11), 881–911. https://doi.org/10.1016/s0145-2126(00)00070-9
Feng, R., Xie, Y., Lai, M., Chen, D. Z., Cao, J., & Wu, J. (2022). AGMI: Attention-guided multi-omics integration for drug response prediction with graph neural networks. arXiv. https://doi.org/10.48550/arXiv.2112.08366
Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, … Bray, F. (2024). Global cancer observatory: Cancer today. International Agency for Research on Cancer. https://gco.iarc.who.int/today/
Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., … Futreal, P. A. (2017). COSMIC: Somatic cancer genetics at high-resolution. Nucleic Acids Research, 45(D1), D777–D783. https://doi.org/10.1093/nar/gkw1121
Freshour, S. L., Kiwala, S., Cotto, K. C., Coffman, A. C., McMichael, J. F., Song, J. J., Griffith, M., Griffith, O. L., & Wagner, A. H. (2021). The Drug–Gene Interaction Database (DGIdb) 4.0: Mining the druggable genome. Nucleic Acids Research, 49(D1), D1144–D1151. https://doi.org/10.1093/nar/gkaa1084)
Friedman, J. H. (1999). Greedy function approximation: A gradient boosting machine. Technical Report, Stanford University.
Ganesh, K., & Massagué, J. (2021). Targeting metastatic cancer. Nature Medicine, 27(1), 34–44. https://doi.org/10.1038/s41591-020-01195-4
Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., … Benes, C. H. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature, 483(7391), 570–575. https://doi.org/10.1038/nature11005
Geeleher, P., Zhang, Z., Wang, F., Gruener, R. F., Nath, A., Morrison, G., … Huang, R. S. (2017). Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. Genome Research, 27(10), 1743–1751. https://doi.org/10.1101/gr.221077.117
Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine

Learning, 63(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1
Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R., … Sellers, W. R. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature, 569(7757), 503–508. https://doi.org/10.1038/s41586-019-1186-3
Gottesman M. M. (2002). Mechanisms of cancer drug resistance. Annual review of medicine, 53, 615–627. https://doi.org/10.1146/annurev.med.53.082901.103929
Gottesman, M. M., Lavi, O., Hall, M. D., & Gillet, J.-P. (2016). Toward a better understanding of the complexity of cancer drug resistance. Annual Review of Pharmacology and Toxicology, 56, 85–102. https://doi.org/10.1146/annurev-pharmtox-010715-103111
Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., … Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA, 316(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216
Ha, S., Park, J., & Jo, K. (2025). Comparative analysis of regression algorithms for drug response prediction using GDSC dataset. BMC Research Notes, 18(Suppl 1), 10. https://doi.org/10.1186/s13104-024-07026-w
Haibe-Kains, B., El-Hachem, N., Birkbak, N. J., Jin, A. C., Beck, A. H., Aerts, H. J., ...Quackenbush, J. (2013). Inconsistency in large pharmacogenomic studies. *Nature*, *504*(7480), 389–393. https://doi.org/10.1038/nature12831
Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. Cell, 144(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013
Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
Holohan, C., Van Schaeybroeck, S., Longley, D. B., & Johnston, P. G. (2013). Cancer drug resistance: An evolving paradigm. Nature Reviews Cancer, 13(10), 714–726. https://doi.org/10.1038/nrc3599
International Cancer Genome Consortium. (2010). International network of cancer genome projects. *Nature,* *464*(7291), 993–998. https://doi.org/10.1038/nature08987
Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., … Garnett, M. J. (2016). A Landscape of Pharmacogenomic Interactions in Cancer. Cell, 166(3), 740–754. https://doi.org/10.1016/j.cell.2016.06.01
Iyer, C. G. S., Languillon, J., Ramanujam, K., Tarabini-Castellani, G., De las Aguas, J. T., Bechelli, L. M., … Sundaresan, T. (1971). WHO co-ordinated short-term double-blind trial with thalidomide in the treatment of acute lepra reactions in male lepromatous patients. Bulletin of the World Health Organization, 45(6), 719–732.
Jain, S., Chouzenoux, E., Kumar, K., & Majumdar, A. (2023). Graph regularized probabilistic matrix factorization for drug–drug interactions prediction. IEEE Journal of Biomedical and Health Informatics, 27(5), 2565–2574. https://doi.org/10.1109/JBHI.2023.3246225
Jain, R. K. (2005). Normalization of tumor vasculature: An emerging concept in antiangiogenic therapy. Science, 307(5706), 58–62. https://doi.org/10.1126/science.1104819
Jensen, M. A., Ferretti, V., Grossman, R. L., & Staudt, L. M. (2017). The NCI

Genomic Data Commons as an engine for precision medicine. *Blood, 130*(4), 453–459. https://doi.org/10.1182/blood-2017-03-735654

Johnson, J. A., Caudle, K. E., Gong, L., Whirl-Carrillo, M., Stein, C. M., Scott, S. A., … Wadelius, M. (2017). Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for Pharmacogenetics-Guided Warfarin Dosing: 2017 Update. Clinical pharmacology and therapeutics, 102(3), 397–404. https://doi.org/10.1002/cpt.668

Jones, L. D., Golan, D., Hanna, S. A., & Ramachandran, M. (2018). Artificial intelligence, machine learning and the evolution of healthcare: A bright future or cause for concern? Bone & Joint Research, 7(4), 223–225. https://doi.org/10.1302/2046-3758.73.BJR-2017-0147.R1

Joo, M., Park, A., Kim, K., Son, W.-J., Lee, H. S., Lim, G., … Kim, J. H. (2019). A deep learning model for cell growth inhibition IC50 prediction and its application for gastric cancer patients. International Journal of Molecular Sciences, 20(24), 6276. https://doi.org/10.3390/ijms20246276

Kang, S., & Hong, S. (2011). Prediction of personalized drugs based on genetic variations provided by DNA sequencing technologies. Genes & Genomics, 33(6), 591–603. https://doi.org/10.1007/s13258-011-0124-z

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, United States.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In Proceedings of ICNN'95 - International Conference on Neural Networks (Vol. 4, pp. 1942–1948). IEEE. https://doi.org/10.1109/ICNN.1995.488968

Komatsu, N., Nakauchi, H., Miwa, A., Ishihara, T., Eguchi, M., Moroi, M., … Yawata, Y. (1991). Establishment and characterization of a human leukemic cell line with megakaryocytic features: dependency on granulocyte-macrophage colony-stimulating factor, interleukin 3, or erythropoietin for growth and survival. Cancer research, 51(1), 341–348.

Kumar, G., Yadav, S., Mukherjee, A., Hassija, V., & Guizani, M. (2024). Recent advances in quantum computing for drug discovery and development. IEEE Access, 12, 64491–64509. https://doi.org/10.1109/ACCESS.2024.3376408

Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N., & Werb, Z. (2018). Tumour heterogeneity and metastasis at single-cell resolution. Nature cell biology, 20(12), 1349–1360. https://doi.org/10.1038/s41556-018-0236-7

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. https://doi.org/10.1109/5.726791

Liu, Y., & Bodmer, W. F. (2006). Analysis of P53 mutations and their expression in 56 colorectal cancer cell lines. Proceedings of the National Academy of Sciences, 103, 976–981. https://doi.org/10.1073/pnas.0509256103

Long, X., Sun, K., Lai, S., Liu, Y., Su, J., Chen, W., … Huang, K. (2025). Artificial intelligence and anti-cancer drugs' response. Acta Pharmaceutica Sinica B. Advance online publication. https://doi.org/10.1016/j.apsb.2025.05.009

Longley, D. B., & Johnston, P. G. (2005). Molecular mechanisms of drug resistance. Journal of Pathology, 205(2), 275–292. https://doi.org/10.1002/path.1706

Lord, C. J., & Ashworth, A. (2016). BRCAness revisited. Nature Reviews Cancer,

16(2), 110–120. https://doi.org/10.1038/nrc.2015.21
Masters J. R. (2000). Human cancer cell lines: fact and fantasy. Nature reviews. Molecular cell biology, 1(3), 233–236. https://doi.org/10.1038/35043102
Mao, Y., Shangguan, D., Huang, Q., Xiao, L., Cao, D., Zhou, H., & Wang, Y.-K. (2025). Emerging artificial intelligence-driven precision therapies in tumor drug resistance: Recent advances, opportunities, and challenges. Molecular Cancer, 24(1), 123. https://doi.org/10.1186/s12943-025-02321-x
Meng, W., Xu, X., Xiao, Z., Gao, L., & Yu, L. (2025). Cancer drug sensitivity prediction based on deep transfer learning. International Journal of Molecular Sciences, 26(6), 2468. https://doi.org/10.3390/ijms26062468
Mergen, B., Çoban, M., Özkan, Ş. S., Başaran, Ö. F., & Özcan, G. (2026). Drug Sensitivity Prediction Using Machine Learning on Integrated COSMIC, DGIdb, and GDSC Data. IEEE Access, 14, 17825-17841
Mergen, B., & Özcan, G. (2025). Makine Öğrenmesi Modelleri ile Kanser Hücre Hattı–İlaç Etkileşim Tahmini. Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi, 25(3), 695-703.
Meyerson, M., Gabriel, S. B., & Getz, G. (2010). The Cancer Genome Atlas: A Comprehensive Analysis of Cancer Genomes. Nature, 472(7341), 120–126. https://doi.org/10.1038/nature09977
Mirabelli, P., Coppola, L., & Salvatore, M. (2019). Cancer cell lines are useful model systems for medical research. Cancers (Basel), 11(8), 1098. https://doi.org/10.3390/cancers11081098
Najgebauer, H., Perron, U., & Iorio, F. (2021). Redefining false discoveries in cancer data analyses. Nature computational science, 1(1), 22–23. https://doi.org/10.1038/s43588-020-00008-5
Naulaerts, S., Dang, C. C., & Ballester, P. J. (2014). Precision and recall oncology: Combining multiple gene mutations for improved identification of drug-sensitive tumours. Oncotarget, 5(0), 97025–97040.
Naveed, S., Husnain, M., Samad, A., Ikram, A., Afreen, H., Gilanie, G., & Alsubaie, N. (2025). Drug efficacy recommendation system of glioblastoma (GBM) using deep learning. IEEE Access, 13, 10398–10411. https://doi.org/10.1109/ACCESS.2024.3514912
Nazarian, R., Shi, H., Wang, Q., Kong, X., Koya, R. C., Lee, H., … Lo, R. S. (2010). Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. Nature, 468(7326), 973–977. https://doi.org/10.1038/nature09626
Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., ... Gray, J. W. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell, 10(6), 515–527. https://doi.org/10.1016/j.ccr.2006.10.008
Olivier, M., Hollstein, M., & Hainaut, P. (2010). TP53 mutations in human cancers: Origins, consequences, and clinical use. Cold Spring Harbor Perspectives in Biology, 2(1), a001008. https://doi.org/10.1101/cshperspect.a001008
Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., … Tyers, M. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Nucleic Acids Research, 49(D1), D529–D541. https://doi.org/10.1093/nar/gkaa1074
Paez, J. G., Jänne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., … Meyerson, M. (2004). EGFR mutations in lung cancer: Correlation with clinical

response to gefitinib therapy. Science, 304(5676), 1497–1500. https://doi.org/10.1126/science.1099314

Park, A., Joo, M., Kim, K., Son, W.-J., Lim, G., Lee, J., … Nam, S. (2022). A comprehensive evaluation of regression-based drug responsiveness prediction models using cell viability inhibitory concentrations (IC50 values). Bioinformatics, 38(10), 2810-2817. https://doi.org/10.1093/bioinformatics/btac177

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In Advances in Neural Information Processing Systems (Vol. 31). Curran Associates, Inc.

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., … Pirmohamed, M. (2019). Drug repurposing: Progress, challenges and recommendations. Nature Reviews Drug Discovery, 18(1), 41–58. https://doi.org/10.1038/nrd.2018.168

Qin, G., Dai, J., Chien, S., Martins, T. J., Loera, B., Nguyen, Q. H., … Becker, P. S. (2024). Mutation Patterns Predict Drug Sensitivity in Acute Myeloid Leukemia. Clinical cancer research : an official journal of the American Association for Cancer Research, 30(12), 2659–2671. https://doi.org/10.1158/1078-0432.CCR-23-1674

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., … Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225. https://arxiv.org/abs/1711.05225

Rauschert, S., Raubenheimer, K., Melton, P. E., & Huang, R. C. (2020). Machine learning and clinical epigenetics: A review of challenges for diagnosis and classification. Clinical Epigenetics, 12(1), 51. https://doi.org/10.1186/s13148-020-00842-4

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., … Hurles, M. E. (2006). Global variation in copy number in the human genome. Nature, 444(7118), 444-454. https://doi.org/10.1038/nature05329

Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B. B., … Chinnaiyan, A. M. (2007). Oncomine 3.0: Genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia, 9*(2), 166–180. https://doi.org/10.1593/neo.07112

Rodgers, P., & Collings, A. (2021). What have we learned? eLife, 10, e75830. https://doi.org/10.7554/eLife.75830

Shah, N. P., Nicoll, J. M., Nagar, B., Gorre, M. E., Paquette, R. L., Kuriyan, J., & Sawyers, C. L. (2007). Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. Cancer Cell, 2(2), 117–125. https://doi.org/10.1016/S1535-6108(02)00096-X

Sharifi-Noghabi, H., Jahangiri-Tazehkand, S., Smirnov, P., Hon, C., Mammoliti, A., Nair, S. K., … Haibe-Kains, B. (2021). Drug sensitivity prediction from cell line-based pharmacogenomics data: Guidelines for developing machine learning models. Briefings in Bioinformatics, 22(6), bbab294. https://doi.org/10.1093/bib/bbab294

Shaker, B., Tran, K. M., Jung, C., & Na, D. (2021). Introduction of advanced methods for structure-based drug discovery. Current Bioinformatics, 16(3), 351–363. https://doi.org/10.2174/1574893615999200703113200

Shaker, B., Ahmad, S., Lee, J., Jung, C., & Na, D. (2021). In silico methods and tools for drug discovery. Computers in Biology and Medicine, 137, 104851. https://doi.org/10.1016/j.compbiomed.2021.104851

Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. Nature Reviews Cancer, 6(10), 813–823. https://doi.org/10.1038/nrc1951

Singh, A., & Settleman, J. (2010). EMT, cancer stem cells and drug resistance: An emerging axis of evil in the war on cancer. Oncogene, 29(34), 4741–4751. https://doi.org/10.1038/onc.2010.215

Song, S., Wu, H., Wang, F., Jiao, J., Xu, L., Wang, H., … Yan, H. (2022). Global research trends and hotspots on glioma stem cells. Frontiers in oncology, 12, 926025. https://doi.org/10.3389/fonc.2022.

Sramkoski, R. M., Pretlow, T. G., Giaconia, J. M., Pretlow, T. P., Schwartz, S., Sy, M. S., … Jacobberger, J. W. (1999). A new human prostate carcinoma cell line, 22R upsilon 1. In Vitro Cell & Developmental Biology - Animal, 35, 403–409. https://doi.org/10.1007/s11626-999-0043-7

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. Nature, 458(7239), 719–724. https://doi.org/10.1038/nature07943

Suda, K., Rivard, C. J., Mitsudomi, T., & Hirsch, F. R. (2017). Overcoming resistance to EGFR tyrosine kinase inhibitors in lung cancer, focusing on non-T790M mechanisms. Expert review of anticancer therapy, 17(9), 779–786. https://doi.org/10.1080/14737140.2017.1355243

Sudo, H., Kodama, H. A., Amagai, Y., Yamamoto, S., & Kasai, S. (1983). In vitro differentiation and calcification in a new clonal osteogenic cell-line derived from newborn mouse calvaria. Journal of Cell Biology, 96, 191–198. https://doi.org/10.1083/jcb.96.1.191

Şentürk, N., Tuncel, G., Doğan, B., Aliyeva, L., Dundar, M. S., Özdemir Sağ, Ş., … Ergören, M. Ç. (2021). BRCA variations risk assessment in breast cancers using different artificial intelligence models. Genes, 12(11), 1774. https://doi.org/10.3390/genes12111774

The Cancer Genome Atlas Research Network. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics, 45*(10), 1113–1120. https://doi.org/10.1038/ng.2764

Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., … Lundberg, E. (2017). A subcellular map of the human proteome. Science, 356(6340), eaal3321. https://doi.org/10.1126/science.aal3321

Topalian, S. L., Hodi, F. S., Brahmer, J. R., Gettinger, S. N., Smith, D. C., McDermott, D. F., … Sznol, M. (2012). Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. The New England journal of medicine, 366(26), 2443–2454. https://doi.org/10.1056/NEJMoa1200690

Vasan, N., Baselga, J., & Hyman, D. M. (2019). A view on drug resistance in cancer. Nature, 575(7782), 299–309. https://doi.org/10.1038/s41586-019-1730-1

Vesteghem, C., Brøndum, R. F., Sønderkær, M., Sommer, M., Schmitz, A., Bødker, J. S., … Bøgsted, M. (2020). Implementing the FAIR Data Principles in precision oncology: Review of supporting initiatives. Briefings in Bioinformatics, 21(3), 936–945. https://doi.org/10.1093/bib/bbz044

Vestias, M.P. 2020. Deep Learning on Edge: Challenges and Trends. In Smart Systems Design, Applications, and Challenges; IGI Global: Hershey, PA, USA, 2020; 23–42

van Schaik R. H. (2008). CYP450 pharmacogenetics for personalizing cancer therapy. Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy, 11(3), 77–98. https://doi.org/10.1016/j.drup.2008.03.002

Volchenboum, S. L., Li, C., Li, S., Attiyeh, E. F., Reynolds, C. P., Maris, J. M., … George, R. E. (2009). Comparison of primary neuroblastoma tumors and derivative early-passage cell lines using genome-wide single nucleotide polymorphism array analysis. Cancer research, 69(10), 4143–4149. https://doi.org/10.1158/0008-5472.CAN-08-3112

Wall, P., & Ideker, T. (2024). Representing mutations for predicting cancer drug response. Bioinformatics, 40(Suppl 1), i160–i168. https://doi.org/10.1093/bioinformatics/btae209

Wan, Z., Sun, X., Li, Y., Chu, T., Hao, X., Cao, Y., & Zhang, P. (2025). Applications of artificial intelligence in drug repurposing. Advanced Science. Advance online publication. https://doi.org/10.1002/advs.202411325

Wang, Y., Fang, J., & Chen, S. (2016). Inferences of drug responses in cancer cells from cancer genomic features and compound chemical and therapeutic properties. Scientific Reports, 6, 32679.

Wang, L., McLeod, H. L., & Weinshilboum, R. M. (2011). Genomics and drug response. New England Journal of Medicine, 364(12), 1144–1153. https://doi.org/10.1056/NEJMra1010600

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association, 58(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Wei, D., Liu, C., Zheng, X., & Li, Y. (2019). Comprehensive anticancer drug response prediction based on a simple cell line–drug complex network model. BMC Bioinformatics, 20(1), 44. https://doi.org/10.1186/s12859-019-2611-1

Weinstein, J. N., Myers, T. G., O'Connor, P. M., Friend, S. H., Fornace, A. J., Jr, Kohn, K. W., … Paull, K. D. (1997). An information-intensive approach to the molecular pharmacology of cancer. Science (New York, N.Y.), 275(5298), 343–349. https://doi.org/10.1126/science.275.5298.343

White, D. L., Saunders, V. A., Dang, P., Engler, J., Zannettino, A. C., Cambareri, A. C., … Hughes, T. P. (2006). OCT-1-mediated influx is a key determinant of the intracellular uptake of imatinib but not nilotinib (AMN107): reduced OCT-1 activity is the cause of low in vitro sensitivity to imatinib. Blood, 108(2), 697–704. https://doi.org/10.1182/blood-2005-11-4687

Willson, J. K., Bittner, G. N., Oberley, T. D., Meisner, L. F., & Weese, J. L. (1987). Cell culture of human colon adenomas and carcinomas. Cancer Research, 47, 2704–2713.

World Health Organization. (2024). Global cancer burden growing, amidst mounting need for services. World Health Organization. https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing-amidst-mounting-need-for-services

Xia, F., Allen, J., Balaprakash, P., Brettin, T., Garcia-Cardona, C., Clyde, A., ... Stevens, R. (2022). A cross-study analysis of drug response prediction in cancer cell lines. Briefings in Bioinformatics, 23(1), bbab356. https://doi.org/10.1093/bib/bbab356

Yang, L., Shi, P., Zhao, G., Xu, J., Peng, W., Zhang, J., … Cui, H. (2020). Targeting

cancer stem cell pathways for cancer therapy. Signal Transduction and Targeted Therapy, 5(1), 8. https://doi.org/10.1038/s41392-020-0110-5

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., … Garnett, M. J. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic acids research, 41(Database issue), D955–D961. https://doi.org/10.1093/nar/gks1111

Zeng, X., Wang, F., Luo, Y., Kang, S.-G., Tang, J., Lightstone, F. C., & Cheng, F. (2022). Deep generative molecular design reshapes drug discovery. Cell Reports Medicine, 3(12), 100794. https://doi.org/10.1016/j.xcrm.2022.100794

Zhang, N., Wang, H., Fang, Y., Wang, J., Zheng, X., & Liu, X. S. (2015). Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. PLoS Computational Biology, 11(9), e1004498. https://doi.org/10.1371/journal.pcbi.1004498

Zhou, Y., Arribas, G. H., Turku, A., Jürgenson, T., Mkrtchian, S., Krebs, K., … Lauschke, V. M. (2021). Rare genetic variability in human drug target genes modulates drug response and can guide precision medicine. Science advances, 7(36), eabi6856. https://doi.org/10.1126/sciadv.abi6856

7. SİMGELER VE KISALTMALAR

AAC	:	area above the dose–response curve
AML	:	akut miyeloid lösemi
AUC	:	Eğri Altındaki Alan
BioGRID	:	Biological General Repository for Interaction Datasets
C	:	ceza parametresi
CCLE	:	Cancer Cell Line Encyclopedia
CNA	:	kopya sayısı değişiklikleri
CNV	:	copy number variations
COSMIC	:	Catalogue of Somatic Mutations in Cancer
CTRP	:	Cancer Therapeutics Response Portal
DGIdb	:	Drug-Gene Interaction database
DL	:	derin öğrenme
DNA	:	deoksiriboz nükleik asit
DSS	:	ilaç duyarlılık skoru
DS-CTR	:	çift katmanlı güçlendirilmiş işbirlikçi konu regresyonu
EHR	:	elektronik sağlık kayıtları
EMT	:	Epitelden mezenkimale geçiş
GDC	:	Genomic Data Commons
GDSC	:	Genomics of Drug Sensitivity in Cancer
HPA	:	The Human Protein Atlas
IC50	:	yarı maksimum inhibitör konsantrasyon
ICGC	:	International Cancer Genome Consortium
KKH	:	Kanser Kök Hücreleri
MAE	:	Mean Absolute Error (Ortalama mutlak hata değeri)
ML	:	Makine öğrenmesi
mRNA	:	haberci (messenger) ribonükleik asit
NCI	:	The National Cancer Institute
NCI-60	:	Ulusal Kanser Enstitüsü-60 (İnsan Tümör Hücre Hatları Taraması)
NLP	:	Natural Language Processing (Doğal Dil İşleme)
PSO	:	parçacık sürüsü optimizasyonu (particle swarm optimization)
R^2	:	determinasyon katsayısı – modelin açıklayıcılık oranı
RMSE	:	Root Mean Squared Error (Kök Ortalama Kare Hatası)
ROC	:	Alıcı Operatör Karakteristiği
SCLC	:	küçük hücreli akciğer kanseri
sigma (σ)	:	RBF çekirdek fonksiyonu genişliği
SNP	:	tek nükleotid varyant

SVM : Support Vector Machines (destek vektör makineleri)
SVR : destek vektör regresyonu
TMB : tumor mutational burden (tümör mutasyon yükü)
TCGA : The Cancer Genome Atlas (Kanser Genom Atlası)
VAF : varyant alel frekansı
WHO : World Health Organization
XAI : Açıklanabilir yapay zeka

8. TEŞEKKÜR

Doktora eğitimim süresince desteğini, bilgisini ve tecrübelerini benimle paylaşan, tezimle ilgili kafamdaki tüm soru işaretlerini gidermek için sorduğum soruları sonsuz sabır ve hoşgörüyle yanıtlayan çok değerli tez danışmanım Doç. Dr. Gıyasettin Özcan'a, eğitimim boyunca engin tecrübeleri ve bilgileriyle gelişmeme katkıda bulunan bölüm Anabilim Dalı Başkanım Prof. Dr. Halil Sağlam'a, değerli bilgilerini öğrencileriyle özveriyle paylaşan Prof. Dr. Şehime Gülsün Temel'e ve herbiri çok değerli Anabilim Dalı Öğretim Üyelerine, engin bilgisi ve çalışma azmiyle her zaman yanımda olan Doç. Dr. Zülfiye Gül'e, çalışmama destek olan sevgili Ömer Faruk Başaran, Meral Çoban, Şenay Sueda Özkan'a, eğitimim sürecinde bana destek olan değerli eşim Kemal Semir Mergen'e, sevgili kızlarım Nil Su Mergen ve Ece Su Mergen'e, sevgili annem Saliha Semra Karakuzu, babam Sinan Karakuzu ve kardeşim Mehmet Görkem Karakuzu'ya yürekten teşekkürlerimi sunuyorum.

9. ÖZGEÇMİŞ

İlk ve orta öğrenimimi babasının görevi dolayısıyla Ankara çevre ilçelerinde, lise öğrenimini ise Bursa Kız Lisesi'nde tamamlamıştır.

1991 yılında İstanbul Üniversitesi Eczacılık Fakültesi'nde başladığı lisans eğitimini 1995 yılında iyi dereceyle bitirmiştir. 2018 yılında Bursa Uludağ Üniversitesi Sağlık Bilimleri Enstitüsü Tıbbi Farmakoloji Anabilim Dalı'nda yüksek lisans öğrenimini tamamlamıştır. 2019 yılında Bursa Uludağ Üniversitesi Sağlık Bilimleri Enstitüsü Translasyonel Tıp Anabilim Dalı'nda doktora eğitimine başlamıştır.

Üniversite eğitimi sonrası ilaç sanayiinde üç yıl Steril Solüsyon Üretim Departmanı'nda üretim şefi olarak görev yapmıştır. İstanbul'da sekiz yıl Maltepe Şifa Eczanesi'nin sahibi ve eczacısı olarak çalışmıştır. 2006–2018 yılları arasında Berfu Eczanesi (Bursa) sahibi ve eczacısı olarak çalışmıştır. 2018 yılından itibaren Sağlık Bakanlığı'nda denetmen eczacı olarak görev yapmaktadır.