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

# YAPAY ÖĞRENME BAZLI HESAPLAMALI MODELLEME İLE GENİŞ ÇAPLI KANSER HÜCRE HATTI İLAÇ YANIT TAHMİNİ

**Umut Onur ÖZCAN**

**Biyoinformatik Programı**

**YÜKSEK LİSANS TEZİ**

**ANKARA**
**2022**

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

# YAPAY ÖĞRENME BAZLI HESAPLAMALI MODELLEME İLE GENİŞ ÇAPLI KANSER HÜCRE HATTI İLAÇ YANIT TAHMİNİ

**Umut Onur ÖZCAN**

**Biyoinformatik Programı**

**YÜKSEK LİSANS TEZİ**

**TEZ DANIŞMANI**

**Doç. Dr. Tunca Doğan**

**ANKARA**
**2022**

**HACETTEPE ÜNİVERSİTESİ**

**SAĞLIK BİLİMLERİ ENSTİTÜSÜ**

**Yapay Öğrenme Bazlı Hesaplamalı Modelleme İle Geniş Çaplı Kanser Hücre Hattı İlaç Yanıt Tahmini**

**Öğrenci: Umut Onur Özcan**

**Danışman: Doç. Dr. Tunca Doğan**

**İkinci Danışman: -**

Bu tez çalışması 07/09/2022 tarihinde jürimiz tarafından "Biyoinformatik Programı" nda yüksek lisans tezi olarak kabul edilmiştir.

| | | |
|---|---|---|
| **Jüri Başkanı:** | *Doç. Dr. Yeşim Aydın Son*<br>*(Orta Doğu Teknik Üniversitesi)* | *(imza)* |
| **Tez Danışmanı:** | *Doç. Dr. Tunca Doğan*<br>*(Hacettepe Üniversitesi)* | *(imza)* |
| **Üye:** | *Doç. Dr. Ceren Sucularcı*<br>*(Hacettepe Üniversitesi)* | *(imza)* |

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukarıdaki jüri tarafından uygun bulunmuştur.

*Prof. Dr. Müge YEMİŞÇİ ÖZKAN*

**Enstitü Müdürü**

## YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan "Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge" kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

o Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. (1)

o Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ... ay ertelenmiştir. (2)

o Tezimle ilgili gizlilik kararı verilmiştir. (3)

… / … / …

Umut Onur Özcan

---

1"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"

(1) Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.

(2) Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.

(3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir *. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.
Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir.

* Tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu tarafından karar verilir

## ETİK BEYAN

Bu çalışmadaki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi, görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu, kullandığım verilerde herhangi bir tahrifat yapmadığımı, yararlandığım kaynaklara bilimsel normlara uygun olarak atıfta bulunduğumu, tezimin kaynak gösterilen durumlar dışında özgün olduğunu, Doç. Dr. Tunca DOĞAN danışmanlığında tarafımdan üretildiğini ve Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Tez Yazım Yönergesine göre yazıldığını beyan ederim.

Umut Onur Özcan

# TEŞEKKÜR

Bir ferdi olarak mutluluk duyduğum, emeklerimin onun varlığına bir armağan olması amacıyla atıldığım bu görevde varlığımı ve heyecanımı diri tutan, aydınlanmamızın en büyük eseri olan eşsiz Türk Devrimimize ve onun önderi Gazi Mustafa Kemal Atatürk'e,

Hayatımın bir diğer dönüm noktasında bana destek olan aileme, Muhittin Gerilecek'e, hocam Doç. Dr. Tunca Doğan'a, TÜSEB projesinde beraber çalıştığımız Navid Mohammadvand'a, projemizdeki in vitro doğrulama deneylerini gerçekleştirip DeepResponse'nin potansiyelinin ortaya çıkartılmasında büyük emekleri olan Doç. Dr. Deniz Cansen Kahraman'a ve Etkin Akar'a,

Değerli arkadaşım Tudor'a, çalışmalarım sırasında eserlerine sayısız ziyarette bulunduğum Ferry Corsten'a, Paul van Dyk'a, Robert Miles'a, Lee Burridge'e, Dwig'e, Artbat'a, Tijs Vervest'e, Dj Precision'a, Hiver & Hammer'a, Fergie'ye, Four Tet'e, Underworld'a, Paul Oakenfold'a, Eelke Kleijn'e, Orkidea'ya, Ken Ishii'ye, Way Out West'e, Steve Helstrip'e,

Teşekkürlerimi sunarım.

Bu çalışma, Türkiye Sağlık Enstitüleri Başkanlığı (TÜSEB) tarafından desteklenmiştir. Proje No: 3912

## ÖZET

**Özcan, U.O., Yapay Öğrenme Bazlı Hesaplamalı Modelleme İle Geniş Çaplı Kanser Hücre Hattı İlaç Yanıt Tahmini, Hacettepe Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoinformatik Programı, Yüksek Lisans Tezi, Ankara 2022.** Her hasta için özel olarak en iyi tedavi seçeneğini değerlendirmek, hassas tıbbın ana hedefidir. Özellikle farklı kanser tiplerinde, aynı tanıya sahip hastalar genetik heterojenite nedeniyle uygulanan tedaviye farklı seviyelerde duyarlılık gösterebilirler. İlaç yanıtlarını (duyarlılığını) önceden tahmin ederek ilaç geliştirme süreçleri için gereken süreden tasarruf etmek ve etkisiz ilaçların uygulanmasını önlemek amacıyla, hastaların genetik özelliklerini kullanan hesaplamalı yaklaşımlar geliştirilmiştir. Bu tez çalışmasında, kanser hücrelerinin ilaç yanıtlarını tahmin eden makine öğrenmesi tabanlı bir sistem olan DeepResponse-RF önerilmiştir. DeepResponse-RF, büyük ölçekli profilleme/tarama projelerinden elde edilen ve her biri ayrı bir tümörü temsil eden farklı kanser hücre hatlarının gen ekspresyonu, mutasyon, kopya sayısı varyasyonu ve metilasyon profillerini, ilaçların moleküler özellikleriyle birlikte kullanmaktadır ve rastgele orman algoritması aracılığıyla, tümörün multi-omik özellikleri ile uygulanan ilaca duyarlılığı arasındaki ilişkiyi yapay olarak öğrenmektedir. Performans sonuçları, DeepResponse-RF'nin kanser hücrelerinin ilaç duyarlılığını başarılı bir şekilde tahmin ettiğini ve özellikle multi-omik yönün öğrenme sürecine fayda sağladığını ve tek omik tabanlı duruma kıyasla daha iyi performansa yol açtığını gösterdi. DeepResponse-RF, daha ileri seviyede geliştirme aşamalarının uygulanması sonrasında, yeni ilaç adaylarının erken aşamada keşfedilmesi ve mevcut olanların dirençli tümörlere karşı yeniden konumlandırılması için kullanılabilir.

**Anahtar Kelimeler:** Hassas Tıp, Makine öğrenmesi, Farmakogenomik test, İlaç Yeniden konumlandırma, Neoplazmlar

"Derin Öğrenme Bazlı Farmakogenomik Modelleme ile Geniş Çaplı Kanser Hücre Hattı İlaç Yanıt Tahmini", TÜSEB Proje kodu : 3912

## ABSTRACT

**Özcan, U.O., Comprehensive Cancer Cell Line Drug Response Prediction by Machine Learning Based Computational Modelling, Hacettepe University, Graduate School of Health Sciences, Department of Bioinformatics, Master's Degree Thesis, Ankara 2022.** Assessing the best treatment option specifically for each patient is the main goal of precision medicine. The patients with the same diagnosis may display varying sensitivity to the applied treatment due to genetic heterogeneity, especially in cancers. With the aim of predicting drug response in advance, to save valuable time and prevent the administration of ineffective drugs, computational approaches that utilize genetic features of patients have been developed. In this thesis study, DeepResponse-RF is proposed, which is a machine learning-based system that predicts drug responses (sensitivity) of cancer cells. DeepResponse-RF utilizes gene expression, mutation, copy number variation and methylation profiles of different cancer cell-lines (each representing an individual tumor) obtained from large-scale profiling/screening projects, together with drugs' molecular features at the input level and process them via the random forest algorithm, to learn the relationship between multi-omics features of the tumor and its sensitivity to the drug administered. Performance results indicated DeepResponse-RF successfully predicts drug sensitivity of cancer cells, and especially the multi-omics aspect benefited the learning process and yielded better performance compared to the single-omic-based state-of-the-art. With further development, DeepResponse-RF can be used for early stage discovery of new drug candidates and for repurposing the existing ones against resistant tumors.

**Keywords:** Precision Medicine, Machine Learning, Pharmacogenomic Testing, Drug Repositioning, Neoplasms

"Derin Öğrenme Bazlı Farmakogenomik Modelleme ile Geniş Çaplı Kanser Hücre Hattı İlaç Yanıt Tahmini", TÜSEB Proje kodu : 3912

# İÇİNDEKİLER

## SİMGELER VE KISALTMALAR

| | |
|---|---|
| **2B** | İki boyutlu |
| **aCGH** | Array Comparative Genomic Hybridization |
| **ABD** | Amerika Birleşik Devletleri |
| **API** | Application Programming Interface |
| **ATP** | Adenozin trifosfat |
| **AUC** | Area Under Curve |
| **BO** | Birleştirilmiş omikler |
| **ÇAA** | Çapraz Alan Analizi |
| **CCLE** | Cancer Cell Line Encyclopedia |
| **ÇD** | Çapraz doğrulama |
| **CGP** | Cancer Genome Project |
| **DNA** | Deoksiribonükleik asit |
| **EC50** | Maksimum ilaç etkisinin yarısına denk gelen konsantrasyon değeri |
| **ECFP** | Extended Connectivity Fingerprint |
| **Emax** | Maksimum ilaç etkisi durumundaki konsantrasyon değeri |
| **GDSC** | Genomics of Drug Sensitivity of Cancer |
| **Gİ** | Gen ifade |
| **GI50** | Growth inhibition 50 |
| **HHÖTB** | Hücre hattı özdeşlik temelli bölümlendirme |
| **Hp** | Hiperparametre |
| **HSK** | Hepatosellüler kanser |
| **IC50** | Yarı maksimal inhibitör konsantrasyonu |
| **İÖTB** | İlaç özdeşlik temelli bölümlendirme |
| **KSD** | Kopya sayısı değişimi |
| **L1000** | Landmark 1000 |
| **LC50** | Lethal concentration 50 |
| **MAE** | Mean absolute error |
| **MET** | Metilasyon |
| **$Mg^{+2}$** | Magnezyum (iki değerlikli) |
| **MGH** | Massachusetts General Hospital |
| **miRNA** | Mikro RNA |
| **MPİ** | Moleküler parmak izi |
| **MSE** | Mean squared error |
| **MUT** | Mutasyon |
| **NCI-60** | National Cancer Institute - 60 |
| **NSC** | National Service Center |
| **ODTÜ** | Orta Doğu Teknik Üniversitesi |

| | |
|---|---|
| **PCC** | Pearson correlation coefficient |
| **PDO** | Patient-derived organoid |
| **PDX** | Patient-derived xenograft |
| **pGI50** | GI50 değerinin negatif logaritma sonucu |
| **pIC50** | IC50 değerinin negatif logaritma sonucu |
| **PICNIC** | Predicting integral copy numbers in cancer |
| **PUG** | Power User Gateway |
| **$R^2$** | Belirleme katsayısı |
| **REST** | Representational State Transfer |
| **RF** | Random forest |
| **RMSE** | Root mean squared error |
| **RNA** | Ribonükleik asit |
| **RNA-seq** | RNA sequencing |
| **RB** | Rastgele özdeşlik temelli bölümlendirme |
| **RRBS** | Reduced bisulfite sequencing |
| **RT-CES** | Real-time cell-impedance sensing |
| **SCC** | Spearman's correlation coefficient |
| **SMILES** | Simplified molecular input line entry |
| **SNP** | Single nucleotide polymorphism |
| **SRB** | Sulforhodamine B |
| **SYP** | Sinyal yolağı profilleri |
| **TCGA** | The Cancer Genome Atlas |
| **TGI** | Total growth inhibition |
| **TÜSEB** | Türkiye Sağlık Enstitüleri Başkanlığı |
| **VİAA** | Veri içi alan analizi |
| **Vb.** | Ve benzeri |
| **WTS** | Wellcome Trust Sanger Institute |

# ŞEKİLLER

# TABLOLAR

# 1. GİRİŞ

Kanser türlerine yönelik etkin ilaçların keşfinin büyük zorluklar ve yüksek maliyetler barındırması, araştırmacıları yeni yaklaşımlarla çözüm getirme yoluna itmiştir. Teknolojideki ilerlemelerin biyolojik veriyi daha erişilebilir ve ucuz kılmasıyla beraber kompleks hastalıkların tedavisi için hassas tıp uygulamalarının önü açılmıştır. Bu amaçla, tümör dokusunun modellenebilmesi için, yaygın olarak hızlı sonuç alınabilmesi ve az maliyetli oluşu nedeniyle hücre hatları kullanılmaktadır. Farmakogenomik paneller ise ilaçların hücre hatları üzerindeki etkisini incelemeyi ve hücrelere ait moleküler özelliklerin ortaya çıkarılmasını hedeflemektedir. Hesaplamalı yaklaşımlar ile moleküler özellik verisi içindeki ilişkiler belirlenebilmekte ve deneysel verisi olmayan benzer özelliğe sahip başka hücre hatları için ilaç yanıtı tahmini yapılabilmektedir. Böylelikle, ileride bu yaklaşımla hastalar için daha doğru tedavi seçenekleri daha az iş gücü ve maliyet ile belirlenebilecektir.

## 1.1. Problem Tanımı

İlaç geliştirme aşamalarının yüksek maliyetli ve zaman alıcı olmasının yanında yeni ilaç onaylanma oranları gittikçe düşmektedir. Şimdiye kadar geliştirilen ilaç yanıtı tahmin yöntemlerinin performans düşüklüğü, çoklu omik veri tiplerini ve ilaç tanımlayıcıları içermemesi en büyük sorunlar olarak karşımıza çıkmaktadır.

## 1.2. Varsayım

Hücre hattı ilaç yanıtının birden fazla değişkene dayanması, hücrenin moleküler özelliklerinin bu yanıtların oluşmasında önemli rol oynamasına bağlı olarak bu özellikleri temsil eden çoklu omik veri tiplerinin kullanılmasıyla ilaç yanıtının modellenebilmesi araştırmamızda temel aldığımız varsayımlardır.

### 1.3. Amaç

Bu tez çalışmasında, hücre hatlarının moleküler özelliklerini temsil edici çoklu omik verinin ve ilaç tanımlayıcı parmak izi dizilerinin beraber kullanılmasıyla yüksek başarıma sahip ilaç yanıtı tahmini modellerini kapsayan DeepResponse-RF yönteminin geliştirilmesi amaçlanmıştır.

### 1.4. Gerçekleştirilen Hedefler

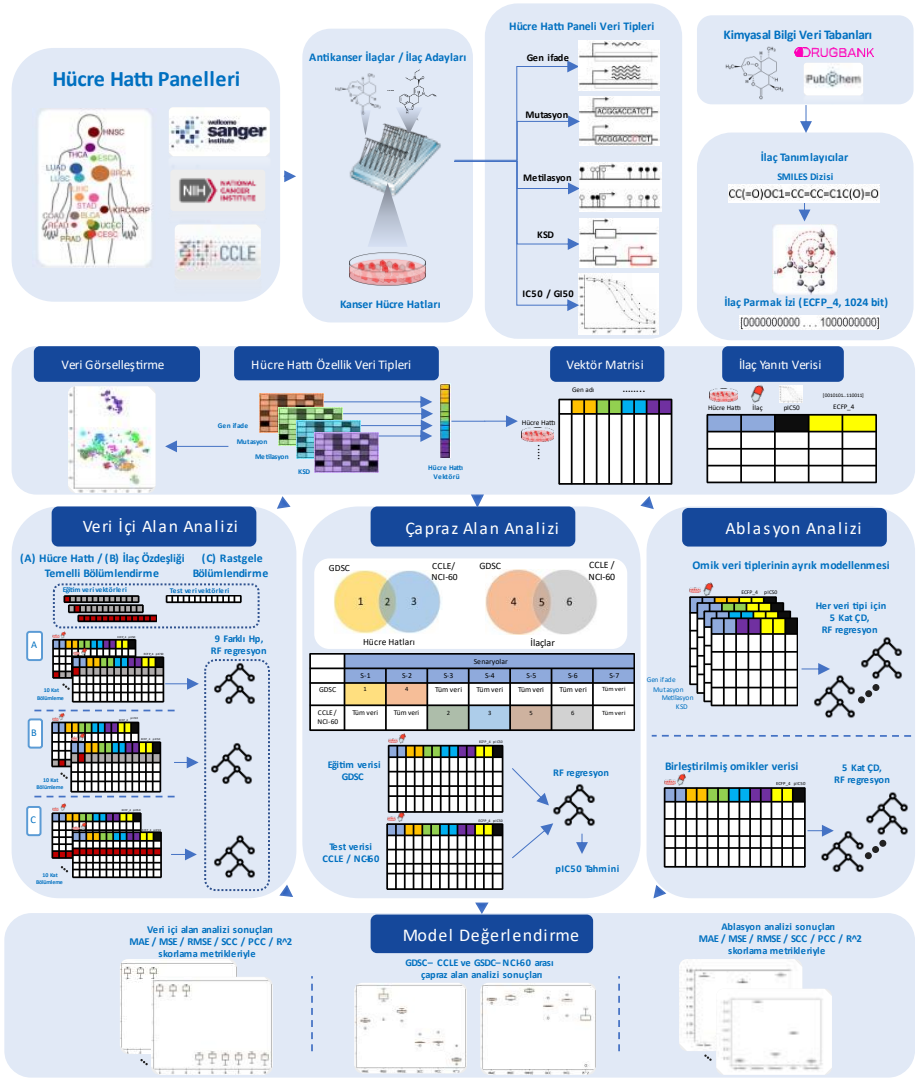
Yukarıda belirtilen amacın gerçekleştirilmesi için ilk aşama olarak koyulan hedefler aşağıda verilmiştir. İlk olarak, hücre hattı panellerinden (GDSC, CCLE, NCI-60) veri tipleri elde edilmesi hedeflenmiştir. Ön işleme aşamalarından geçirilen omik veri tiplerinin son aşamada birleştirilmesi amaçlanmıştır. İlaçlar için ise SMILES dizi karşılıkları veri tabanlarından elde edilip bu dizilerden ilaç parmak izi değerleri üretilmesi planlanmıştır.

Veri tiplerinin birleştirilmiş halde kullanılmasının tahmin performansına etkisini ölçmek için GDSC verisi üzerinden ablasyon analizi yapılması planlanmıştır. Bu analiz tipinde, her omik veri ve birleştirilmiş omikler verisi için ayrı ayrı tahmin modelleri oluşturulması hedeflenmiştir. Sonrasında, veri içi alan analizinde GDSC verisi doku temelli değerlendirilip üç bölümlendirme yöntemiyle model eğitim setinde olmayan hücre hatları, ilaçlar veya hücre hattı-ilaç çiftleri için tahminler oluşturulması amaçlanmıştır. Devamında ise çapraz alan analiziyle model eğitim (GDSC) ve test setinin (CCLE veya NCI-60) farklı panellere ait olduğu iki durum için (GDSC - CCLE ve GDSC - NCI-60) hücre hattı veya ilaç ortaklıklarını göz önüne alan yedi farklı modelleme senaryosunun kurgulanması hedeflenmiştir. Benzer analiz tiplerini kullanan yöntemler literatürden bulunarak DeepResponse-RF sonuçlarıyla karşılaştırılması planlanmıştır.

DeepResponse-RF ile GDSC'nin sindirim sistemi doku verisi için üretilen tahminlerin bazıları *in vitro* doğrulamalarıyla desteklenmesi hedeflenmiştir.

Hepatosellüler kanser (HSK) hücre hatları üzerinde yapılan *in vitro* deneyleri, TÜSEB projesinde ortağımız olan ODTÜ KanSiL Lab tarafından gerçekleştirilecek şekilde planlanmıştır.

Son olarak, DrugBank veri tabanında bulunan ilaçlar kullanılarak GDSC'nin sindirim sistemi doku verisi ile eğitilen DeepResponse-RF ile, dokuda bulunan 99 hücre hattı ve parmak izi çıkarılan DrugBank ilaçlarının oluşturduğu tüm hücre hattı – ilaç çiftleri için tahminler oluşturulması hedeflenmiştir. DeepResponse-RF'de gerçekleştirilen genel akışının gösterimi Şekil 1.1'de sunulmaktadır.




**Şekil 1.1.** DeepResponse-RF'de gerçekleştirilen genel iş akışı.

## 2. GENEL BİLGİLER

İnsanlarda ilaç yanıtının hücrede görülen genetik varyasyonlarla ilişkisinin ortaya çıkarılması farmakogenetik alanının oluşmasından bugüne kadarki ana hedefi olmuştur. Gelişen sekanslama teknolojileriyle beraber ilaç yanıtına etki eden belirteçlerin ortaya çıkarılmasıyla farmakogenetik alanı (farmakogenetik ve farmakogenomik alanlarının arasındaki fark kesin olmadığından günümüzdeki çalışmalar bu tez çalışması kapsamında sadece farmakogenomik alanı göz önüne alınarak incelenmiştir) büyük bir değişim geçirip tanımlayıcı bilimden öngörücü bilim anlayışına geçmiştir. 1990'lı yıllarda yapılan çalışmalarla bir hastalığın aynı seviyesinde olan hastaların uygulanan tedaviye farklı genetik değişimlere sahip oldukları için farklı yanıtlar geliştirebileceği öngörülmüştür (1).

150 Binden fazla araştırma makalesinde önerilen biyobelirteçlerin içinden sadece 100'den azının klinikte kullanılabildiği raporlanmıştır (2). Birçok biyobelirteç günümüze kadar fenotip değişkenliği sorunu dolayısıyla gelememiş ve klinikte uygulama alanı bulamamıştır. Bu nedenle, tutarlı fenotip ve biyobelirteç tanımlamalarının standartlaştırılması yoluna gidilmiştir. Böylelikle, hem yan etki reaksiyonları azaltılmış olacak hem de standartlaştırılan biyobelirteçler daha güvenli ve doğru şekilde kullanılabilecektir (2). Bununla beraber, biyobelirteçlerin daha detaylı verilere dayanarak oluşturulması ve ilaç yanıtının daha doğru tahmini için farklı moleküler verilerden de yararlanıp hasta veya tümör fenotipine özel tedavi seçenekleri oluşturulması beklenmektedir (3).

Genetik varyasyonlar ve ilaç molekülleri arasındaki ilişkiyi bulma farmakogenomik alanının asıl amaçlarından biridir. Bu doğrultuda yapılan araştırmalarla, daha az deney sayısı ve yüksek doğrulukla ilaç etkinliği tespit edilerek hasta yararı, iş gücü, klinik uygulama, zaman, ekonomi açısından da olumlu katkı sağlanmaktadır. Kullanılacak ilacın etkinlik ve güvenlik sorunlarının birçoğu bu alandaki yaklaşımlarla çözülme potansiyeli olduğu için araştırmacıların dikkati bu yöne çevrilmiştir (4). Yüksek çıktılı sekanslama

teknolojisiyle elde edilen büyük hacimli biyolojik veri tiplerinin birleştirilip hastalığı anlamlı şekilde temsil ederek kullanılması farmakogenomiğin uygulama alanına uygun düşmektedir (5).

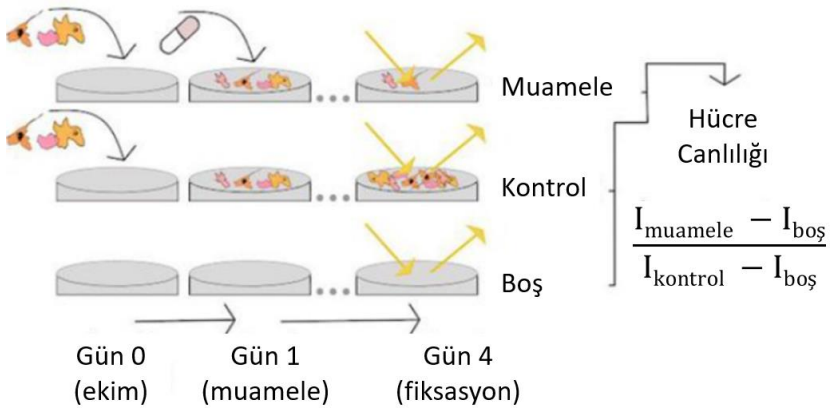
İlaçla tedavi yöntemlerinin hassas tıp yaklaşımlarıyla beraber iyileştirilebileceğinin düşünülmesine rağmen son yıllarda geliştirilen yeni ilaçların onaylanma hızı ve oranlarının düşmesi, yüksek etkinlik seviyelerinde olmaması araştırmacıları başka yollar aramaya sevk etmiştir (2,6,7). Yeni ve etkinliği yüksek ilaçların geliştirilmesini farmakogenetik ve farmakogenomik yaklaşımlarla destekleyen translasyonel araştırmalar, keşif çalışmalarına ivme kazandırmıştır. Böylece, hastaların farklılıklarını tanıyıp onların profillerinde daha etkili olabilecek tedavilerin bulunabilmesini kolaylaştıracak yöntemlerin geliştirilmesine olanak sağlanmıştır (6). Ancak, literatürde bulunan tahmin modellerinin geniş ölçekte uygulanabilir olmayışı önemli bir problem olarak görülmektedir. Hesaplamalı yaklaşıma dayanan bu modellerde çoklu omik verinin etkili şekilde kullanılamaması sebebiyle tahmin performanslarının düşük seviyelerde kaldığı saptanmıştır.

Dünya çapında hastalık bazlı ölümlerin nedeni ve uzun yaşam beklentisinin önündeki en büyük engel sayılan kanserin 2011-2015 yılları arasında aldığı can sayısının önemli derecede düşüş gösterdiği görülmüştür. Araştırmacılar bu olumlu tabloyu erken teşhis ve daha etkin şekilde uygulanan tedavi yaklaşımlarıyla ilişkilendirmektedir. Bu bağlamda, en iyi tedavinin belirlenebilmesinde hesaplamalı modellerin kısıtlı veri ve algoritmik seviyedeki karşılaştığı zorluklara rağmen, hastalığın ilaca karşı yanıtının tahmini konusunda büyük öneme sahip olduğu düşünülmektedir. Yakın zamanda geliştirilen ve ileri seviye algoritma mimarileri kullanan ilaç yanıtı tahmini modelleri, hastaların hayatta kalma şanslarını artırmaya destek olan araçlar olarak karşımıza çıkmaktadır (8).

### 2.1. İlaç Yanıtı ve Tahmini

*In vitro* tarama deneylerinde izlenen genel strateji, bir ilacın bir hücre hattı kültürü üzerinde denenmesi üzerinedir. Hücrelerin ilaca karşı gösterdikleri hassasiyet, hücre ölümü veya gelişim durdurma gibi olgularla sonuçlanabilir. İlaç yanıtı, bu gösterilen hassasiyetin deneyin amacına göre kullanılan metriklerle kantitatif olarak ifadesidir.

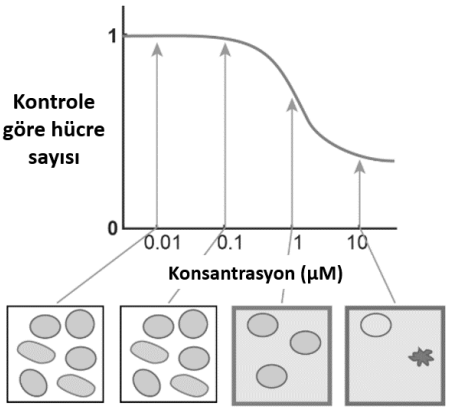
İlaç yanıtı analizlerinde uygulanan prosedürlerde benzer yöntemler takip edilmektedir. Genel olarak, ilaçla muamele edilen hücre kültürü ile kontrol hücre kültüründeki hücre sayısı karşılaştırılır. İlaç eklenen kültür için birden fazla örnek oluşturularak bu örnekler üzerine aynı ilacın farklı konsantrasyonları eklenir. İlacın uygulama süreci bittikten sonra (analiz prosedürüne göre 24 ve 72 saat arasında değişkenlik gösteren süreler uygulanabilir), her bir örnekteki hücre sayısı hesaplanır (9). Yukarıda değinilen sürece ait genel akış, Şekil 2.1.'de belirtilmiştir.




**Şekil 2.1.** İlaç yanıtında analizlerinde hücre canlılığı ölçümü için uygulanan genel akış ve hücre canlılığı kantifikasyonu için kullanılan formül (7).

Sonuçların bir grafik üzerinde gösterilmesi, değerlerin anlamlandırılması açısından önemlidir. Uygulanan konsantrasyon değerleri artan şekilde apsis eksenine yerleştirilir. Ordinat ekseninde 1 (örnekte başlangıçta bulunan tüm hücreler, % 100) ve 0 (örnekte ilaç uygulanmasından sonra canlılık gösteren hücrenin kalmadığı durum, % 0) arasındaki aralık gösterilir. İlaç uygulama süresi sonucunda örneklerdeki hücre sayısı

hesaplandıktan sonra grafiğe aktarılır. Şekil 2.2.'de, bahsedilen grafiğin ters bir s eğrisi oluşturan sigmoid fonksiyon örneği belirtilmiştir.



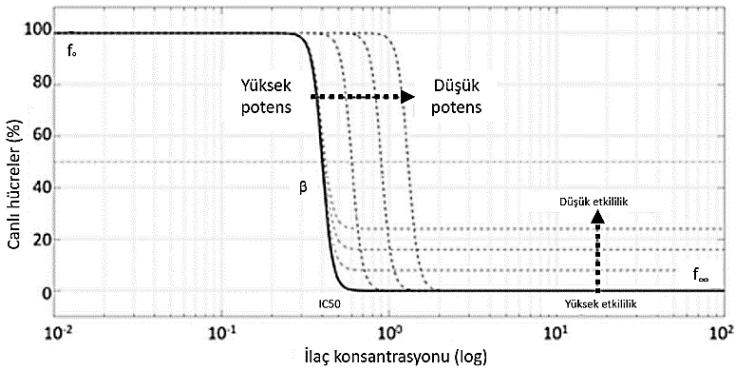

**Şekil 2.2.** Hücre hattı üzerinde uygulanan farklı konsantrasyon değerleri için hücre sayısındaki değişimin gösterilmesi.

İlaç yanıtı analizlerinde, sonuçların grafiğe aktarılmasında genellikle dört değişkenli bir lineer olmayan lojistik regresyon denklemi kullanılır (Formül 2.1.). Bu denklem, oksijenin hemoglobin yapısına bağlanmasının kantifikasyonu için 1910'da A. V. Hill tarafından geliştirilmiş olan denklemin değiştirilmiş halidir (10,11). Formül 2.1.'de f(x), x konsantrasyonundaki ilaç yanıtını ifade etmektedir. $f_0$ parametresi, uygulanan düşük ilaç konsantrasyonlarında gözlenen asimptotu ifade eder. $f_\infty$, yüksek ilaç konsantrasyonlarında görülen asimptotu ifade eder ve ilaç etkililiğinin bir göstergesidir. $f_\infty$ (bazı kaynaklarda $E_{max}$ olarak da belirtilmektedir (9)) değerinin grafikte düşük seviyelerde olması, ilacın etkililiğinin yüksek olduğu anlamına gelir. IC50 (yarı maksimal inhibitör konsantrasyonu, half maximal inhibitory concentration), örnekteki hücre popülasyonunun yarısının üzerinde istenilen etkinin görüldüğü konsantrasyondur. Aynı zamanda, IC50, yukarıda bahsedilen iki asimptotun arasında kalan eğrinin orta noktasıdır ve farklı terimlerle de ifade edilebilmektedir. ß, Hill katsayısı veya eğri eğimi olarak da ifade edilir ve ilaç yanıtındaki değişkenliği gösterir. Ek olarak, eğim ne kadar dik ise ilaç yanıtı da o ölçüde homojendir denilebilir. Sigmoid eğrinin asimptot çizgileri değişmeden sola doğru kayması durumunda ise ilaç yanıtı daha düşük konsantrasyon değerleri alır.

Sonuç olarak, ilaca ait potens artmış olur ve ilaç daha az doz ile istenen etkinliği gösterir (Şekil 2.3.) (11,12). Formül 2.2. ise f∞'in 0 olduğu durumlarda formülün genel kullanım formatını göstermektedir.

$$f(x) = f_\infty + \frac{f_0 - f_\infty}{1 + \left(\frac{x}{\mathrm{IC}_{50}}\right)^{-\beta}} \tag{2.1.}$$

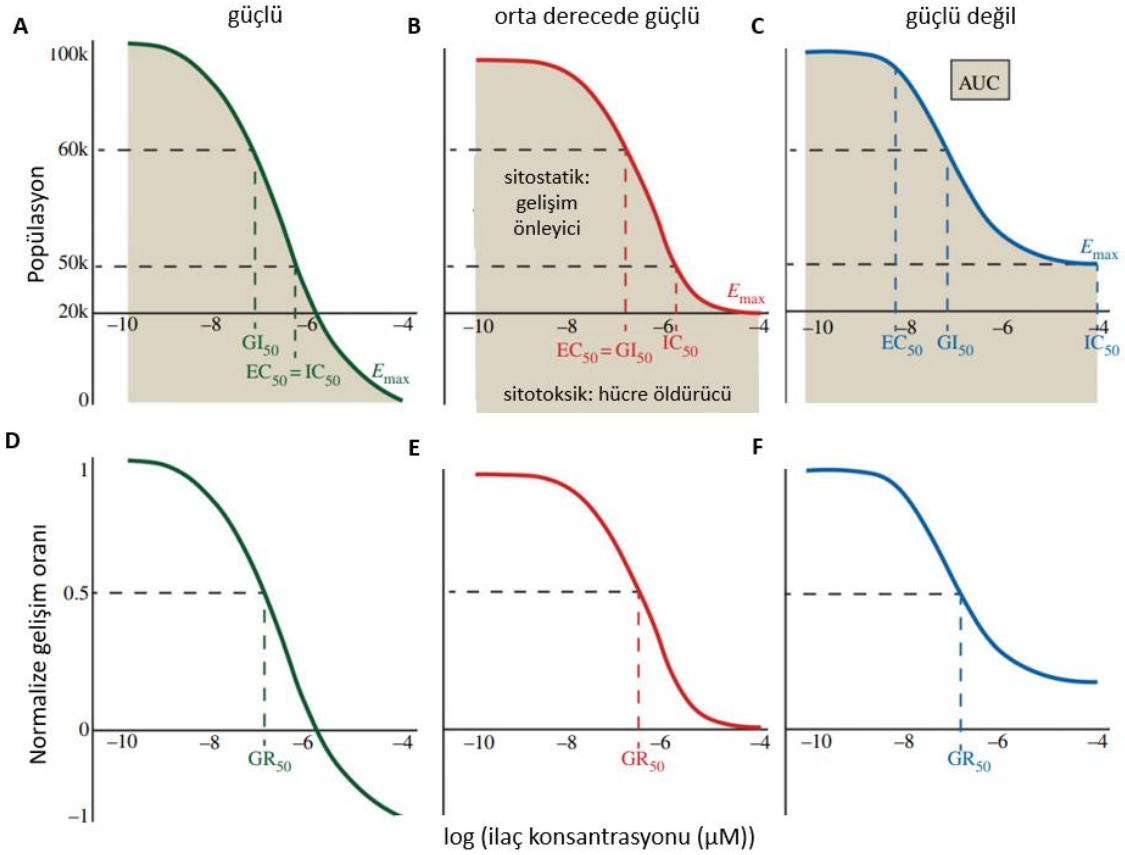
$$f(\gamma_0) = \frac{100}{1 + \left(\frac{\gamma_0}{\mathrm{IC}_{50}}\right)^{-\beta}} \tag{2.2.}$$




**Şekil 2.3.** Hücre hattı örneği üzerine uygulanan ilacın potensi ve etkililiğinin ilaç yanıtı eğri grafikleriyle örneklendirilmesi (12).

İlaç yanıtı analizinde ölçülen değerler hücre canlılığı haricinde hücre gelişim oranı için de olabilir. Buna bağlı olarak, Formül 2.1.'de yapılan hesaplama sonucu oluşan sigmoid eğri üzerinden yapılan çıkarımda, örnekteki hücrelerin yarısında (analiz amacına göre hücrelerin ölümünü başlatacak veya gelişimini engelleyecek şekilde) etkili olacak ilaç konsantrasyonunun belirlenmesi sağlanır. Yukarıda bahsedilen farklı amaçlar için ilaç konsantrasyonlarının hesaplanmasında amaca özel metrikler kullanılmaktadır. Bunlardan tez çalışmamızda kullandığımız iki metrikten biri, örnekteki canlı hücre

popülasyonunu (başlangıçtaki canlı hücre sayısına göre) yarı yarıya azaltan ilaç konsantrasyonudur ve IC50 olarak ifade edilmektedir. Kullanılan ikinci metrik ise, gelişim engelleyici 50 (growth inhibition 50, GI50) adı verilen, örnekteki tüm hücrelerin gelişimini % 50 oranında azaltan ilaç konsantrasyonudur. Diğer yaygın olarak kullanılan metrikler Tablo 2.1.'de ve Şekil 2.4.'te özetlenmiştir.




**Şekil 2.4.** İlaç yanıtı metriklerinin farklı ilaç etkinlik durumları için grafik üzerinde örneklendirilmesi (9). Uygulanan ilacın etki durumu üç kategoride sunulmuştur; güçlü (A ve D); orta derecede güçlü (B ve E); güçlü değil (C ve F). Düşey eksende yer alan ikili grafik grupları, sırasıyla popülasyon ve normalize gelişim oranına denk gelen aynı durumları ifade etmektedir.

**Tablo 2.1.** İlaç yanıtı analizlerinde kullanılan metriklere ait tanımlamalar ve onlara ait konsantrasyonların hesaplanma formülleri (9).

| Metrik | Tanım | Formülleştirme |
|---|---|---|
| IC50 | Popülasyonu kontrol değerinin yarısına kadar azaltan konsantrasyon | (popülasyon) = 0.5 x $y_{kontrol}$ |
| GI50 | Toplam hücre gelişimini % 50'ye kadar azaltan konsantrasyon | (popülasyon) = 0.5 x ($y_{kontrol} - y_0$) + $y_0$ |
| GR50 | Hücre gelişim oranını % 50'ye kadar azaltan konsantrasyon | (popülasyon) = $2^{\log2(yGI50/y_0)/(yGI50/y_0)} - 1$ |
| EC50 | $E_{max}$ ile belirlenen etkinin yarısına denk gelen konsantrasyon | (popülasyon) = 0.5 x ($y_{kontrol} - E_{max}$) + $E_{max}$ |

$E_{max}$, ilaç etkisiyle mümkün olan maksimum düzeydeki gelişim engellemesi veya hücre ölümü.

### 2.2. Kanser Hücre Hatlarında İlaç Yanıtı

Klinik öncesi modellerin geliştirilmesi sırasında genel olarak kullanılan ilaç etkinliğinin denendiği üç yöntem gösterilmektedir. Bunlar, insan tümöründen elde edilen hücre hatlarıyla oluşturulan kültürler, farelerde oluşturulan ksenograftlar ve genetiği değiştirilmiş fare modelleri. Bu üç yöntemin birbirine üstünlüğü fizyolojik temsiliyet açısından araştırmacılar arasında tartışma konusudur (13).

Günümüzde, insan kanser hücre hattı kültürlerinin ilaç geliştirme aşamalarında ve geniş kapsamlı farmakogenomik taramalarda yaygın olarak kullanıldığı görülmektedir. Buna sebep olarak, fare modellerinin yüksek çıktılı ilaç tarama çalışmalarına uygun olmaması, kanseri ayırt edici özelliklerden genomik heterojenitenin olmaması (özellikle genetiği değiştirilmiş farelerde), sonuçların insanlarda da aynı olmaması, zaman alıcı prosedürler, büyük ekonomik yük ve iş yükü gösterilebilir. Belirtilen bu dezavantajların

çoğunun kanser hücre hatları için geçerli olmaması nedeniyle farmasötik ve biyoteknoloji endüstrisinde hatların geniş kullanım alanı bulduğu belirtilmektedir (13–15).

Hayvan tümörlerine bir alternatif olarak gösterilen kanser hücre hatları, kendi fizyolojik ortamlarıyla aynı olmayan ve uzun dönem pasaj kültürü içinde bulunmaktadırlar. Buna rağmen, elde edildikleri primer tümördeki genetik yapı büyük ölçüde korunduğu için ilaç yanıtı ve genetik özellikler arasında klinik açıdan anlamlı korelasyonlar bulunduğu yapılan çalışmalarla belirlenmiştir (13,14). Kanser hücre hatlarının antikanser ilaç testlerinde kullanılmasının yanı sıra genetik, epigenetik, proteomik, sinyal yolakları ve apoptoz ile ilgili çalışmalarda da genişçe yer bulması bu alanlarda güvenilir bir araç olduğunu göstermektedir (14).

Kanser hücre hatları deney tekrar edilebilirliği, görece diğer yöntemlere göre ucuz olması, genetik heterojenite, kısa zamanda kolayca büyütülüp ilaç molekülleriyle yapılan testlere sonuç verme gibi avantajları sayesinde yüksek çıktılı ilaç tarama paneli çalışmalarına da temel oluşturmaktadır. Bu çalışmalarda, hücre hattı ilaç yanıtlarının kantitatif sonuçları, sekanslama metotlarıyla elde edilen hücreyi genomik açıdan temsil edebilecek çeşitli omik veri tipleri ile desteklenmektedir (14,16–18). Tarama platformlarının bu veri tiplerini herkesin erişebileceği halde kendi veri tabanlarında bulundurması, ilaç yanıtı tahmini modelleme ve ilaç yanıtına etki eden moleküler belirteçlerin belirlenmesi gibi konular üzerine çalışan araştırmacılar için büyük kolaylık sağlamaktadır.

### 2.3 Omik Veriyle İlaç Yanıtı Tahmini

Gün geçtikçe ilerleyip gelişen sekanslama teknolojilerinin artık daha ucuz maliyetli olmaları hastalık çeşitli hastalıklara dair moleküler düzeydeki değişimleri tespit edebilmemizi kolaylaştırmıştır. Kompleks bir hastalık olan kanserde de tümörün ilaçlara ne düzeyde yanıt göstereceğini tahmin edebilmek için omik verileri temel alan yaklaşımlar geliştirilmiştir. Bu yaklaşımı benimseyen çalışmalarda hücre içindeki

değişimleri ifade edebilecek şekilde epigenomik, genomik, transkriptomik, proteomik ve metabolomik gibi tekniklere ait veri tiplerinin kullanılması, ilaç yanıtı olgusunu bir sistem anlayışı içinde araştırılmasına yol açmıştır (3,19).

Kanserli hücrelerin genetik yapılarında meydana gelen gen silinmesi, eklenmesi, translokasyonu ve tek nükleotid polimorfizmi gibi varyasyona sebep olan sapmaların modellenebilmesi, kanser gelişimi ve ilaç geliştirme gibi çalışma alanlarında primer tümör yapısını taklit edebilme açısından oldukça yararlı olacaktır (20,21). Kanser hücre hatları ve bunların üzerinde ilaç moleküllerinin denenmesi üzerine kurgulanan geniş kapsamlı tarama deneylerinde de omik veri tiplerinin ilaçların antikanser etkisinin belirlenebilmesinde oldukça büyük öneme sahip olduğu görülmektedir. Geniş kapsamlı olan bu ölçümlerin yapılabilmesi için başlatılan Kanser Genom Atlası (The Cancer Genome Atlas, TCGA), Kanserde İlaç Hassasiyet Genomiği (Genomics of Drug Sensitivity in Cancer, GDSC), Kanser Hücre Hattı Ansiklopedisi (Cancer Cell Line Encyclopedia, CCLE) ve Ulusal Kanser Enstitüsü – 60 (National Cancer Institute – 60, NCI-60) gibi uluslararası işbirlikleri veya ulusal çaptaki çalışma gruplarının çabalarıyla oluşturdukları projelerde, çoklu omik veri tiplerinin hücre hatları moleküler özelliklerini belirlenmesinde kullanılan ana araç olduğu görülmektedir (17,22–24).

Omik veri tipleri araştırmacılara hem kanser biyolojisine dair hem de kanser tedavisinde uygulanacak yaklaşımın seçilmesi için bir içgörü kazandırmaktadır. Bununla beraber, elde edilen verinin hücrenin mekanistik modellerinde ve makine öğrenmesine dayanan ilişkisel ve korelatif tahminler üreten modellerde kullanılmaya uygun olacağı belirtilmektedir (25).

### 2.3.1. Gen İfade Verisi ve İlaç Yanıtı Tahminindeki Önemi

İlaç yanıtı tahmini modelleme yaklaşımlarında hücre hatlarına ait özelliklerin temsili için gen ifade omik veri tipi uzun zamandır yaygın olarak kullanılmaktadır (26). Mikrodizi veya RNA sekanslama (RNA-Seq) analizleri, verinin elde edilmesinde deneyler

için mali yük getirse de, gen ifade verisinin tahmin modellerinde performansı önemli derecede artıracak bir bilgi kaynağı olduğu pek çok çalışma tarafından raporlanmıştır (27–29).

İlaç yanıtı üzerinde transkript değişimlerinin de etkisinin olduğu düşünülmektedir. RNA-seq yönteminden daha eski olan mikrodizi temelli gen ifade profillemeleri, birçok kanser ve hastalık tedavilerinin karar aşamalarında yaygın kullanım alanı bulmuştur. Ancak, klasik mikrodizi yöntemlerinin çoğu transkript değişimlerini yakalayamamaktadır. Tek nükleotid düzeyinde daha yüksek hassasiyet sunan RNA-Seq yöntemi ise hücrenin tüm transkriptlerini belirleyebilmektedir. Bu nedenle RNA-Seq, bahsedilen sorunun üstesinden gelmek için büyük potansiyel taşımaktadır (30).

Hücrede protein kodlayan binlerce gen olduğundan, gen ifade veri yapısı da sonuç olarak büyük boyutlara ulaşabilmektedir. Tüm genlerin bir arada değerlendirilmesi model performansını olumsuz etkileyeceğinden, genellikle fazlalık olarak görülen genlerin veriden çıkarılması yolu izlenmektedir. Bu amaçla, L1000 (978 gen içeren liste) (31) gibi kanser hücresine ait işaret gen listelerinin kullanılmasıyla ilaç yanıtı tahmin edilen kanser hattı hücrelerinin temsili daha az genle sağlanabilmektedir (27).

### 2.3.2 Mutasyon Verisi ve İlaç Yanıtı Tahminindeki Önemi

Genomik düzeyde meydana gelen değişimlerle bazı ilaçların oluşturabileceği yanıtlar arasında belirli düzeyde bir korelasyon bulunmaktadır. Buradan hareketle, tedavilerde kullanılan ilaçlara karşı gelişen yanıtın tahmininde hastaya ait genetik varyasyonların da yararlı olabileceği düşünülmektedir (32,33). Böylece, hastaya özel olacak şekilde ilaç, doz belirlenmesi kolaylaştırılacak ve yan etki reaksiyonları düşürülebilecektir (32).

Bir ilacın hücre üzerinde etkili olabilmesi için hücre içinde bazı süreçlerden geçip hedeflendiği yapılara bağlanarak işlevini yerine getirmesi beklenir. Hücredeki ilaç

metabolize edici enzimlerin, ilaç taşıyıcı proteinlerin veya hedef proteinlerin yapılarında değişiklikler meydana geldiği durumlarda ilaç yanıtı önemli ölçüde etkilenecektir. Bahsedilen proteinleri kodlayan genetik sekanslarda oluşacak değişimler üründeki yapı değişikliklerine veya protein sentezlenmesinin durmasına neden olabilmektedir (32–34).

Tek nükleotid polimorfizmleri (single nucleotide polymorphism, SNP) somatik hücrelerde oluşan bu genetik varyasyonların ana sebeplerinden birisidir. 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 değişim şekliyle, protein yapısına eklenen farklı bir amino asit ile proteinin fonksiyonunda varyasyonlar (yüksek aktivite veya inaktivasyon gibi) beklenebilir. Bunun yanında, DNA sekansındaki değişimler amino asit ifade eden kodonun bir durdurucu kodona dönüşmesine neden olarak proteinin normalden hızlı şekilde bozunuma uğramasına veya inaktif olmasına sebebiyet verebilir. Sessiz SNP'ler, protein kodlayan sekanslardaki mutasyon durumlarındaki değişimin proteinde işlevsel olarak herhangi bir etki oluşturmamaktadırlar (32).

2010'lu yıllardan günümüze kadarki zaman diliminde farklı çalışma gruplarının yayımladığı çalışmalarda, mutasyon verisinin ilaç yanıtı tahminini pozitif yönde etkilediği gösterilmiştir. Aynı amaçtaki analizlerde bu veri tipinin göz önüne alınması önerilmiştir (33–35). Buradan hareketle, hücre hattı panelleri de genetik değişikliklerin ilaç yanıtını öngörücü özelliğinden faydalanmak için onkogenler ve tümör baskılayıcı genler başta olmak üzere belirteç özelliği gösterecek somatik mutasyonlara ait veri tipini de sunmaktadırlar. Ayrıca, ilaç yeniden konumlandırma amacı için kullanımda olan bazı mutasyon belirteçleri, panellere ait veri ve genetik taramalarının ilaç hedef genleri üzerinde birbirini desteklemeleri sonucuyla ortaya çıkmıştır (35).

### 2.3.3 Metilasyon Verisi ve İlaç Yanıtı Tahminindeki Önemi

İlaç yanıtındaki değişimin köken aldığı genetik heterojeniteye yalnızca DNA sekanslarındaki değişim etki etmemektedir. DNA sekanslarındaki değişimler ile açıklanamayan fenotipik farklılıklara profillenebilen epigenetik faktörlerin de katkı sunduğu bilinmektedir (36–38).

Epigenetik, çevre ve genom arasındaki bağlantısallığı ve yeni nesillere aktarılabilen, ancak genetik olarak kodlanmayan özellikleri incelemektedir. DNA (sekans değişimleri harici) ve histon üzerindeki değişiklikler, kromatin oluşumu, nükleozomlar gibi konular, epigenetiğin çalışma alanına dahildir. Bu değişimlerden birisi olan DNA metilasyonunda, metil parçaları ard arda gelen sitozin ve guanin (CpG) bölgelerindeki sitozinlere DNA metiltransferaz tarafından eklenir. Oluşan metilsitozinler ya gen ifadesini engeller ya da ilgili protein yapıları yardımıyla kapalı kromatin yapısına geçişi başlatarak transkripsiyon faktörlerini engeller (37,39). Çoğu tümörde görülen şekliyle, CpG adacıklarındaki (CpG'ler açısından zengin olan bölgeler, CpG islands, CGI) hipermetilasyon promotor bölgesinde oluştuğunda gen inaktivasyonuna neden olmaktadır (38–40).

Hücredeki pek çok düzenleyici mekanizmanın kontrolünde önemli yeri olan epigenetik faktörler, tümör hücrelerinin oluşumunda da ayırt edici özelliklerden bazılarını oluşturmakta ve onların hayatta kalmak için avantajlarını artırmaktadır (36,37,39). DNA'da meydana gelen anormal seviyedeki metilasyonlar, hidroksimetilasyonlar, hipometilasyonlar gibi epigenomik değişimler kanser hücrelerinde de görülmektedir. Bu değişiklikler, hücredeki düzensizliği artırıp transkripsiyon, hücre döngüsü, gelişme ve çoğalma mekanizmalarını direkt veya dolaylı olarak etkilemektedir. DNA tamiri ve tümör baskılayıcı gibi görevleri olan genlerin baskılanması sonucunda tümör hücrelerinin ilaca karşı yanıtları da farklılaşacaktır. Dahası, metillenerek baskılanan genlere sahip hücrelerin özellikleri, uygulanan ilaç

tedavisinin seçici etkisiyle sonraki nesildeki popülasyona da aktarılabildiğinden ilaç direnci oluşması da mümkündür (39,41). Tümör hücrelerindeki bu belirli değişimler aynı zamanda onları normal doku hücrelerinden ayıracak ipuçlarını da bize göstermektedir. Bu anlamda, metilasyon değişimleri içerdiği değerli bilgilerden dolayı ilaç temelli tedavilerde iyi bir belirteç seçeneği olarak karşımıza çıkmaktadır (36).

Bazı epigenetik faktörleri kodlayan genler ilaç yanıtıyla ilişkilendirildiği için, bu bölgelerin metillenmesi gen ifade düzeylerini değiştirerek ilaç yanıtına etki ettiğine dair kanıtlar yakın zamandaki çalışmalarla daha da artmaktadır (36,41). DNA metilasyonlarının, kanser hücre hatlarındaki ilaç yanıtlarıyla korelasyon gösterdiği raporlandığı için ilişkilerin ileri seviyede araştırılması da önerilmektedir (41,42). Shen ve ark. (36) NCI-60 panel verisi üzerinde yaptıkları sistematik metilasyon ve demetilasyon çalışmasıyla ilaç yanıtının tahmini kolaylaştıracak ve yararlı olabilecek metilasyon belirteçlerinin bir listesini sunmuşlardır (14).

### 2.3.4 Kopya Sayısı Değişimi Verisi ve İlaç Yanıtı Tahminindeki Önemi

Her bir kromozom üzerinde bir çift DNA kopyası bulunduğundan gen bölgelerinin de iki kopya olarak bulunduğu varsayılır. Ancak, İnsan Genom Projesi bu sayının ikiden az veya çok olduğu bazı aleller üzerinden raporlanmıştır. DNA sekanslarında görülen bu değişimler kopya sayısı değişimi (KSD) olarak adlandırılmaktadır. Bahsedilen bu değişimlerdeki sekans uzunluğu, eklenme, çıkarılma, ters çevrilme, çoğaltılma ve farklı rekombinasyon gibi işlemlere bağlı olarak 1 kilobazdan birkaç megabaza kadar farklılık gösterebilmektedir (43).

Gen sekanslarında oluşan değişimlere bağlı olarak, gen ifadesinin düzeyini etkileyen ana nedenlerden birisi de KSD'dir. Ancak, bu değişimler gen çevresinde bulunan sekanslarda oluştuğunda gen ifadesi etkilenmeyebilir (44). İlaç metabolize edici enzimlere ait genlerde oluşacak bu tip değişiklikler, hücrenin ilaçlara karşı yanıtını büyük

ölçüde etkileyecektir. Bu durum, tedavide kullanılacak ilacın göstereceği yarar ve toksisite üzerinde farklılıklara yol açacaktır (45).

Şimdiye kadarki süreçte yeterince göz önüne alınmasa da, yakın zamandaki çalışmalar KSD'nin hastalıklardaki rolüne ve ilaç yanıtıyla ilişkisine dair yeni içgörüler sunduğundan alandaki araştırmacıların ilgisini çekmektedir (44,46). KSD'nin, Crohn's hastalığı ve sinir sistemi hastalıkları, otoimmün bozukluklar gibi kompleks hastalıklar üzerindeki katkısı yapılan araştırmalar sonucu belirlenmiştir (44,45). Gen bazında örnek vermek gerekirse, CYP2D6 geninin kopya sayısının artışıyla aktivitesinin yükselmesi sonucu, kişi ölümcül opioid zehirlenmesine daha yatkın olacağı belirtilmiştir. CYP2A6'da (ilaç metabolize edici enzimleri üreten bir gen) da KSD görülebilmektedir. Artan CYP2A6 gen aktivitesiyle, nikotin bağımlılığı ve tütün ilişkili kanserlere yönelik risk arttığı raporlanmıştır (45).

KSD verisinin elde edilebilmesi için geliştirilen teknolojiler ve onları destekleyici algoritmaların ortaya çıkmasıyla, bu veri tipi ilaç yanıtı araştırmalarında da kullanılabilir hale gelmiştir. Gamazon ve ark. (45)'nın araştırmaları ile KSD'nin ilaç yanıtındaki katkısı gösterilmiştir. Çalışmada, belirli KSD bölgelerinin SNP etiketlemeleriyle (etiketten bağımsız ilişkiler de raporlanmıştır) beraber ilaç yanıtıyla ilişkili olduğu ortaya çıkarılmıştır.

### 2.4. Literatürdeki ilaç yanıtı tahmini modellemesi yapan çalışmalar

İlaç yanıtı tahmini problemi için geliştirilen yöntem sayısının fazlalığı, yöntemlerin kullandığı veri tiplerinin ve modelleme yaklaşımlarının farklılığından ileri gelmektedir. Hücre hatları özelliklerini temsil eden omik veri tipleri açısından bazı modeller tek; bazıları ise çoklu omik veri kullanımını tercih etmişlerdir. Önceden ilaç benzerliği aramalarında destek verici olarak kullanılan ilaç tanımlayıcı veri tipinin tahmin modellerinde kullanımı bir tartışma konusu olmuştur. Bu nedenle, bazı yöntemler ilaç parmak izlerini kullanmamayı tercih ederken, bazıları da kullanmış veya modeli bu veriye

uygun halde tutmuştur (47). Bu bölümde incelenecek olan yöntemler, tahmin modeli oluşturmada kullanılan veri tipi farklılıkları göz önüne alınarak seçilmiştir.

Hücre hattı panellerinin gelişmesiyle, klinik öncesi deneylerde birçok kanser hücre hattı denenebilir hale gelmiştir. Ancak, aynı durum insan üzerinde panellerde kullanılan yüksek sayıdaki ilaçları denemek mümkün olmadığı için hastadan alınan tümör verisinde geçerli değildir. Huang ve ark. (48)'nın yaptığı çalışmada, hücre hattı panel verisi üzerinden hastadaki tümörün verebileceği yanıtın tahmini için bir makine öğrenmesi metodu geliştirilmiştir. Klinik öncesi veri için gen ifade veri tipi ve ilaç yanıtı sonuçları GDSC'den; hastadan alınan tümör örneklerine ait veri ise Kanser Genom Atlası'ndan (The Cancer Genome Atlas, TCGA) elde edilmiştir. TG-LASSO adı verilen model, ilaç yanıtı tahmin sonucunu IC50 değeri ile verebilecek; hastaların ilaca karşı hassasiyet veya direnç gösterenler olarak ayırabilecek şekilde tasarlanmıştır. Performans değerlendirmeleri çapraz veri seti doğrulama ve tekrarlı rastgele alt örnekleme çapraz doğrulama ile yapılmıştır.

Hibrit enterpolasyon ağırlıklı ortak filtreleme (hybrid interpolation weighted collaborative filtering, HIWCF) yöntemi bir öneri sistemi (recommender system) olarak geliştirilmiştir. Öneri sistemlerindeki kullanıcı ve öğe dizisi çifti arasındaki modellemeye benzer olarak, sırasıyla hücre hattı ve ilaç çifti kullanılmıştır. Modelin girdi verisi olarak kullanılan gen ifade veri tipi GDSC ve CCLE'den alınmıştır. İlaç tanımlayıcı veri tipi olarak ise ilaç parmak izi kullanılmıştır. Modelin değerlendirilmesi 10 kat çapraz doğrulama yöntemiyle yapılmıştır. İlaç yanıtı değerleri IC50 ve aktivite bölgesi (activity area) metrikleri cinsinden verilmiştir (49).

Niepel ve ark. (50)'nın yaptığı çalışmada bazal ve ligand ile uyarılmış sinyal yolaklarına ait profiller ve mutasyon durumları değerlendirilerek ilaç yanıtı tahmini yapılmıştır. Kısmi en küçük kareler regresyon (partial least squares regression, PLSR) yöntemi temel alınarak kurgulanan model, tahminlerini GI50 değeri verecek şekilde

tasarlanmıştır. Çalışmada yalnızca NCI-ICBP43 adlı hücre hattına ait veri üzerinde çalışılmıştır. Modelin değerlendirilme aşamalarında tek-çıkışlı çapraz doğrulama (leave-one-out cross validation, LOOCV) yöntemi tercih edilmiştir.

Miranda ve ark. (38)'nın yaptığı araştırmada, özellikle hücre hatlarının metilasyon profilleri üzerinde durularak bu veri tipinin ilaç yanıtı tahminindeki önemi belirlenmeye çalışılmıştır. Toplamda 5 farklı sınıflandırma algoritması ve dört farklı regresyon algoritması tahmin sonuçlarının karşılaştırılması için uygulamada kullanılmıştır. 987 Hücre hattına ait metilasyon verisi GDSC'den elde edilerek modelde girdi verisi olarak kullanılmıştır. Sınıflandırma ve regresyon modellerinden elde edilen tahminler IC50 değeri, sırasıyla ayrık ve sürekli formda oluşturulmuştur. Model değerlendirmelerinde iç içe (nested) çapraz doğrulama yöntemi tercih edilmiştir. Karşılaştırmalar sonucu sınıflandırma algoritmalarının daha üstün performans gösterdiği raporlanmıştır. Diğer yandan, TCGA'dan alınan tümör verisiyle hastaların ilaçlara karşı yanıtı tahmin edilmeye çalışılsa da güvenilir tahmin sonuçları alınamadığı belirtilmiştir. Shen ve ark. (36)'nın 2007'de NCI-60 metilasyon verisini kullanarak 30 binden fazla ilaç için yaptığı çalışmadan farklı olarak burada daha az ilaç kullanılarak modeller oluşturulmuştur.

Gamazon ve ark. (45)'nın 2011 yılında yayımladığı çalışmada, DNA'daki yapısal varyantların farmakolojik fenotipler üzerindeki etkisi araştırılmıştır. Çalışma için geliştirilen yöntemde, genomdaki KSD'ler bulunarak ilaç yanıtına etkisi ortaya çıkarılmıştır. Yaygın olarak kullanılan hücre hattı panellerinden hariç olarak, hücre hatları HapMap phase II CEU çalışmasına ait örneklerden elde edilmiştir. Bu hücre hatlarının bazı antikanser ilaçlarına karşı ilaç yanıtı IC50 metriğiyle ölçülmüştür. Yapılan ilişki analizleriyle, hem SNP etiketli olan hem SNP etiketinden bağımsız olan KSD'lerin ilaç yanıtıyla bağlantılı olduğu lineer regresyon modelleriyle gösterilerek raporlanmıştır.

Yukarıda bahsedilen tek omik veri kullanan yöntemlerin yanı sıra, farklı algoritmik teknikler ve birden çok omik veri tipinin bir arada kullanıldığı yaklaşımlar da

geliştirilmiştir. Bu yaklaşımlardan biri 2013 yılında Menden ve ark. (51) tarafından uygulanmıştır. Çalışmada geliştirilen makine öğrenmesi modelleri (derin sinir ağları ve rastgele orman) üzerinde hücre özellik veri tipleri (mutasyon, KSD, mikro uydu) ve ilaç tanımlayıcılar eğitim verisi olarak kullanılmıştır. Omik veri tipleri için GDSC; ilaç tanımlayıcı için basitleştirilmiş moleküler giriş satır giriş sistemi (*simplified molecular-input line-entry system*, SMILES) veri tipi tercih edilmiştir. Modeller, 8 kat çapraz doğrulama yapılarak kurulup IC50 metriğiyle ilaç yanıtı tahminleri verecek şekilde tasarlanmıştır.

İlaç yanıtı tahmini modellerinde gen ifade verisinin gösterdiği performansın daha iyileştirilebilmesi ve elde edilen sonuçların klinik kullanıma aktarılabilirliğinin araştırılması amacıyla 2019 yılında Sharifi-Noghabi ve ark. (52) çoklu-omik geç birleştirimi (multi-omics late integration, MOLI) adlı yöntemi geliştirmişlerdir. İlaç yanıtı tahminini bir sınıflandırma problemi olarak ele alan MOLI yönteminde, GDSC, TCGA, ve PDX Encyclopedia platformlarından elde edilen veri tipleri farklı deneysel tasarımlar için kullanılmıştır. Yöntemin isminde de yer alan geç birleştirme tekniğiyle gen ifade, mutasyon, KSD veri tipleri birer derin sinir ağıyla farklı akışlarda işlendikten sonra çıkarılan özellikleri birleştirilir. Üçlü kayıp (triplet loss) tekniği ile birbirine yakın örnekler arasındaki uzaklıklar düşürülüp daha benzer hale getirilir. İkili çapraz-entropi kaybı (binary cross-entropy loss) tekniğiyle de hasta örnekleri yanıt veren veya yanıt vermeyen olarak sınıflanır. Diğer bir deneysel tasarımda ise, ön işlemden geçirilen çoklu omik veri tipleri (hasta verisi veya hastadan alınan ksenograft verisi), yanıtı incelenecek olan ilaç için tahmin değeri (0-1 arası) oluşturmada kullanılmıştır.

Çoklu omik profilleri ilaç yanıtı tahmini için kullanan başka bir yöntem olan DeepCDR, 2020 yılında Liu ve ark. (53) tarafından geliştirilmiştir. Genomik profillere ek olarak ilaç yapı bilgisini de göz önüne alan DeepCDR yöntemi hibrit çizge evrişimsel ağ mimarisi temelinde oluşturulmuştur. Çalışmada, GDSC, CCLE, TCGA veri tabanlarından alınan omik veri ve hasta profillerinden ve PubChem kütüphanesinde bulunan ilaç

tanımlayıcı yapısal bilgi dosyalarından yararlanılmıştır. İlaç ve çoklu omik veri tipleri (gen ifade, mutasyon, metilasyon) farklı akışlarda ön işlemden geçirilmiştir. İlaç tanımlayıcı veri için tekdüze çizge evrişimsel ağ kullanılarak yüksek-düzey geç temsil (high-level latent representation) oluşturulmuştur. Diğer yandan, omik verilerin her biri için ayrı alt ağ yapıları kullanılarak profillerin birleştirimi yapılmıştır. İlaç ve omik veri temsillerinin birleştirilmesiyle oluşan veri üzerinden regresyon analizi yapılarak IC50 metriğiyle ilaç yanıtı tahmini oluşturulur.

# 3. GEREÇ VE YÖNTEM

Bu tez çalışması kapsamında GDSC, CCLE, NCI-60 farmakogenomik taramalarına ait hücre hatlarının bazı moleküler özelliklerini temsil eden veri tipleri ve ilaç yanıtı verisi değerlendirilmiştir. Bu bölümde, belirtilen panellere dair genel bilgiler ve veri içeriklerinin ayrıntılı şekilde incelendiği kısımlar sunulmuştur.

Farmakogenomik alanının ortaya çıkmasından bu yana, alandaki araştırmaların hedefinde belirli genelleyici konular bulunmaktadır. Bunlar, hastaların gruplandırılıp tedaviden elde edilecek yararın artırılabilmesi, ilaçların kanser hastaları üzerindeki yanıtının tedavi öncesinde tahmin edilebilmesi hedefleridir. Bu hedefleri gerçekleştirmek için ise genellikle tümör hücrelerinin genomik özellikleri göz önüne alınmıştır. Tümörlerin profillenmesinde kullanılan teknolojilerin oldukça yaygın olmasından dolayı bu alan artık farmako-omik olarak da adlandırılmaktadır (28,38).

Kanser hakkında bizlere önemli bilgiler sunan ilk geniş kapsamlı çalışmalar temel olarak iki ana hedef gütmektedir. Bu hedeflerden ilki, ilaç moleküllerinin kanser hücre hatları üzerinde sistematik şekilde denenmesidir. Diğeri ise, bu hücre hatlarına ait genetik yapının ilaç yanıtındaki etkisini ortaya çıkarmak için çoklu omik verinin elde edilmesi üzerine kurgulanmıştır. Bu çalışmaların öncüsü sayılan ilk adımlar Amerika Birleşik Devletleri'ndeki (ABD) Ulusal Kanser Enstitüsü'nün (National Cancer Institute, NCI) NCI-60 projesiyle atılmıştır (13,19,28,54).

Dünya çapında çeşitli çalışma grupları, NCI-60 ile yaratılan panel paradigmasını benimseyerek yeni projelere öncülük etmişlerdir. Aşağıda belirilen bu projeler, tümör heterojenitesini daha iyi belirtecek şekilde, yeni hücre hatlarını ve klinikte kullanım potansiyeli olan yeni ilaçları panel yapısına katarak genişlemeye devam etmektedir (18,55).

## 3.1. Veri Tabanları

### 3.1.1. GDSC

GDSC projesi, İngiltere'deki Wellcome Trust Sanger Institute (WTS) ve ABD'deki Massachusetts General Hospital (MGH) arasındaki iş birliğinin ürünü olarak ortaya çıkmıştır. Projenin ilk zamanlarındaki isimlendirmesi Kanser Genom Projesi (Cancer Genome Project, CGP) olsa da ilerleyen süreç içinde değiştirilmiştir (19). Projeye ait internet portalında (www.cancerrxgene.org) tüm veri setleri indirmeye açıktır ve veri tabanında sorgulama yapılabilmektedir. Portal, sorgulama sonuçlarının yorumlanmasını kolaylaştırıcı bir grafik arayüz ile desteklenmektedir (22).

Projeye ait internet portalında belirtilen bilgilendirme notuna göre, 2010 ve 2015 arasında gerçekleştirilen çalışmalar GDSC1; daha yeni olan veri setleri ise GDSC2 etiketi ile sunulmaktadır. GDSC1 ilaç yanıtı verisinin oluşturulmasında Resazurin veya Syto60 testleri yapılmış ve 987 hücre hattı ve 367 ilaç kullanılmıştır. GDSC2'de ise ilaç yanıtı analizi olarak CellTitreGlo testi kullanılmış ve 809 hücre hattı ve 198 ilaç molekülü ile veri noktaları üretilmiştir. Projenin en son versiyonu 2020 yılının Haziran ayında çıkan 8.3 sürümüdür (56,57). Bu tez çalışması kapsamında kullanılan GDSC ilaç yanıtı verisinin düzenlenmesinde, projeye portalında yer alan direktifler uygulanmıştır. Bu direktifler çerçevesinde, aynı hücre hattı – ilaç çifti için hem GDSC1 hem GDSC2 veri noktası bulunduğunda, ekipman ve prosedür iyileştirmeleri olduğu için GDSC2 veri noktaları tercih edilmiştir (57).

Projedeki ilaç yanıtı ölçümünde, GDSC1'de hücre canlılığı analizi için uygulanan prosedürlerde resazurin adlı bileşik kullanılmıştır. Hücre içine alınan resazurin bileşiğinin enzimler yardımıyla redükte edilmesiyle oluşan renk değişimi sonucuna dayanarak canlı hücrenin enzimatik aktivite tayini yapılmıştır (58). İlaç yanıtı için GDSC1'de uygulanan diğer yöntem ise nükleik asit analizi temeline dayanan Syto60 yöntemidir. Aynı zamanda bir kolorimetrik analiz yöntemi olan Syto60, adını Syto60 bileşiğinden almaktadır. Bu

bileşik, canlı hücrelerin nükleik asit yapılarına yapışıp kırmızı renge boyamaktadır. Böylece, mikroskop altındaki canlı hücrelerin miktarını rengin ışıma yoğunluğuna orantılı olarak gösterilmesine yardımcı olmaktadır (28). Projedeki her hücre hattı – ilaç ikilisine ait veri noktası, IC50 ve eğim altında kalan alan (Area Under Curve, AUC) metrikleri ile sunulmuştur. GDSC2 için uygulanan yöntem olan CellTitreGlo, bir Adenozin trifosfat (ATP) temelli analiz çeşididir. Bu hücre canlılık analizi yöntemi, ilaçla muamele edildikten sonra canlılığını kaybetmeye yaklaşan hücrelerin normalden az ATP üretmesi olgusuna dayanmaktadır. Analizde kullanılan lusiferaz enzimleri, canlı hücreler içindeki $Mg^{+2}$ (iki değerlikli Magnezyum) ve ATP yardımıyla lusiferine dönüşüp ışıma yaratmaktadır. Ortaya çıkan ışıma yoğunluğuyla canlı hücrelerin miktarı tayin edilmektedir (28).

### 3.1.2. CCLE

CCLE projesi, 2006 yılında ABD'deki Broad Institute ve Novartis Biyomedikal Araştırma Enstitüleri (Novartis Institutes of Biomedical Research) arasındaki işbirliği sonucunda ortaya çıkan bir projedir (19). Projenin ilk adımları üç faz halinde 2008-2017 arasında yapılmış olup, ilk veri setleri hem projenin kendi internet portalında (https://sites.broadinstitute.org/ccle/datasets ve https://data.broadinstitute.org/ccle/) hem de Kanser Değişkenleri Haritası (Cancer Dependency Map, DepMap) veri tabanında (https://depmap.org/portal/download/) yayımlanmaya başlanmıştır. 2018'den günümüze gelen süreçte, projedeki omik veri tipleri hemen hemen yılın her çeyreğinde yenilenerek DepMap veri tabanına eklenmektedir. Projenin en son versiyon olarak 2022 yılının ikinci çeyreğinde kullanıma açılan veri için 22Q2 kodu kullanılmıştır (59).

CCLE'de omik verileri çıkarılan yaklaşık 1000 hücre hattı olmasına karşın ilaç yanıtı ölçülen toplam 504 hücre hattı bulunmaktadır. Diğer panellere nazaran daha az sayıda denenen ilaç bulunduran CCLE'de, toplam 24 ilaç için ilaç yanıtı analizi yapılmıştır (19).

CCLE'de ilaç yanıtı analizi için GDSC2'de olduğu gibi CellTitreGlo yöntemi tercih edilmiştir. Bu yöntem ile 24 ilaç kullanılarak 504 hücre hattının ilaç yanıtı değerleri ölçülmüştür. Ölçülen ilaç yanıtı değerleri için metrik olarak IC50 kullanılmıştır (60).

CCLE'de ilaç yanıtı verisi, gen ifade, mutasyon, metilasyon ve KSD omik veri tiplerinin yanı sıra tam genom sekanslama, tüm ekzom sekanslama, ters faz protein analizi, metabolomiks, genel kromatin profilleme, gen etkisi analizi (CRISPR yöntemiyle gen nakavtı) gibi incelemelere de yer verilmiştir. Bu analizlere ait veri setleri kamuya açık halde ve kullanıma hazır halde bulunmaktadır (14,23).

### 3.1.3. NCI-60

NCI-60 projesi, ilaç keşfini kolaylaştıran ve hayvan modellerinin yerini alabilecek bir aracın geliştirilmesi amacıyla 1980'li yıllarda başlatılmıştır (14,24). Projenin geliştirilme aşamaları üç adımda gerçekleştirilmiştir. Bu aşamalar, *in vitro* ilaç yanıtı analiz çeşitlerinin araştırılması, panelin geliştirilmesi ve panelde kullanılacak bilgi teknolojisinin oluşturulması başlıkları altında toplanmıştır. Proje sırasında geliştirilen teknolojiler, şu anda devam eden diğer ilaç tarama hedefli projelere örnek oluşturmuştur (13,61). Günümüzde NCI-60 projesi, tümör hücrelerinin gelişim engelleme mekanizmaları üzerine çalışan araştırmacılar için zengin içeriğe sahip bir kaynak haline gelmiştir (24).

NCI-60 panelinde, 60 insan kanser hücre hattı üzerinde denenen 130 binden fazla ilaç molekülü bulunmaktadır. Bunlardan yaklaşık 22 bin ilaç yanıtı verisi kamuya açık haldedir. Panelde ilaç yanıtı analizi için Sulforhodamine B (SRB) adlı kolorimetrik analiz tercih edilmiştir. SRB analizinde kullanılan aminoxanthine boyası, düşük seviye asidik koşullardaki canlı hücrenin protein yapılarında bulunan bazik amino asitlere bağlanıp pembe renkte ışıma vermektedir. Bu ışıma yoğunluğu, analiz örneğindeki canlı hücrelerin miktarıyla orantılıdır (28). İlaç yanıtı değerleri için panelde dört farklı metrik kullanılmıştır. Bunlar, IC50, GI50, toplam gelişim engelleyici konsantrasyon (total growth inhibition, TGI) ve öldürücü konsantrasyon 50 (lethal concentration, LC50) şeklindedir

(62,63). Bu tez çalışmasında genel uyumluluk açısından NCI-60 paneli ilaç yanıtı verisi için GI50 metriğindeki değerlerin kullanılması uygun görülmüştür.

NCI-60 panelinde bulunan eski moleküler veri tipleri, işlenmemiş ve ön işlemden geçirilmiş olarak CellMiner (güncel olan versiyon 2022.1, https://discover.nci.nih.gov/cellminer/datasets.do) bünyesinde bulunmaktadır (64). Veri tiplerinin yeni versiyonları (ön işlemden geçirilmiş halde) ise CellMiner Çapraz Veri tabanı'na (CellMiner cross database, CellminerCDB, güncel olan versiyon 1.5, https://discover.nci.nih.gov/rsconnect/cellminercdb/) eklenmektedir. CellMinerCDB, bu tez kapsamında yararlandığımız paneller dahil olmak üzere; PRISM, CTRP, Achilles, MD Anderson gibi projelere ait hücre hattı özellik veri tiplerinin de kataloglandığı geniş kapsamlı bir veri tabanıdır (17,65). CellMinerCDB'de tez çalışması süresince yararlandığımız NCI-60 paneline ait ilaç yanıtı, gen ifade (mikrodizi log2 yoğunluk değerlerinden hesaplanan z skoru, çoklu platform mikrodizi sonuçlarından elde edilen log2 yoğunluklarının ortalama değerleri, RNA-sekanslama verisi), mutasyon, metilasyon, KSD veri tipleri bulunmaktadır. Veri tabanında bunlara ek olarak, mikroRNA ifade, protein ifade (ters-faz protein ve kütle spektrometresi protein analizi), H3K27ac tarafından işaretlenen aktif güçlendirici (enhancer) / düzenleyici (promoter) sinyaller, H3K4me3 tarafından işaretlenen aktif düzenleyici sinyallere ait veri tipleri de yer almaktadır (54,55,66). Farklı kaynaklardan gelen benzer veri tiplerinin hepsinde örtüşecek şekilde isimlerin tek formatta standart olarak atanması ve veri tiplerindeki yeni güncellemelerin veri tabanına eklenebilmesi gibi avantajlar, CellMinerCDB'yi kullanım kolaylığı açısından da öne çıkarmaktadır (65).

Omik verinin ilaç yanıtı ile ilişkilendirilmesi klinik açıdan maliyeti artırıcı bir yol olarak görünse de hastanın çeşitli omik veri tipleriyle oluşturulan profilinin ilaç yanıtı tahmini için önemli bir kazanım sağlayacağı belirtilmektedir (28). Buradan hareketle, tez çalışması için yukarıda belirtilen üç hücre hattı paneli değerlendirmeye alınmıştır. Bu

hücre hattı panellerinin uygulama ekipmanları ve tercih edilen analizlerin genel içerikleri Tablo 3.1.'de gösterilmiştir.

**Tablo 3.1.** Hücre hattı panellerinin uyguladığı analiz prosedürlere ait içerikler ve tercihlerin karşılaştırılması.

| Panel ismi | Hücre hattı sayısı | İlaç sayısı | İlaç yanıtı analizi türü | Hücre hatlarının ilaçla muamele edilme süresi | İlaç yanıtı metriği | Paneli inceleyen ilk çalışmalar |
|---|---|---|---|---|---|---|
| GDSC | 1126 (GDSC1'de 987, GDSC2'de 809) | 518 (GDSC1'de 367, GDSC2'de 198) | Resazurin veya Syto 60 (GDSC1), CellTiter Glo (GDSC2) | 72 saat | IC50, AUC | (51) |
| CCLE | 504 | 24 | CellTiter Glo | 72 saat | IC50, AUC | (60) |
| NCI-60 | 60 | >50000 | SRB | 48 saat | IC50, GI50, TGI, LC50 | (24) |

### 3.2. Gen ifade veri tipi

Hücre hattı panellerinin gen ifade veri tipini elde etmek için kullandığı araçlar ve kaynak veri içeriği Tablo 3.2.'de gösterilmiştir.

**Tablo 3.2.** Hücre hattı panellerine ait gen ifade veri tipinin içerikleri ve verinin elde edilmesinde kullanılan araçlar.

| Panel adı | Kaynak | Hücre hattı sayısı | Gen sayısı | Veri elde etmede kullanılan araç / metot |
|---|---|---|---|---|
| GDSC | Orijinal | 1018 | 17737 | Affymetrix HG-U219A |
| CCLE | CellMiner | 1088 | 19851 | Affymetrix HG-U133 Plus 2.0 |
| NCI-60 | Orijinal (xai kodlu veri) | 60 | 23059 | Affymetrix HG-U95, HG-U133, HG-U133 Plus 2.0; GeneChip Human Exon 1.0 ST array; Agilent Whole Human Genome Oligo Microarray |

Aşağıdaki tablolarda; GDSC, CCLE ve NCI-60 gen ifade kaynak veri tablolarının genel yapıları örneklendirilerek sunulmaktadır. Bu örneklendirmelerde gen ve hücre

hatlarına ait bölümler, satır ve sütun sayıları yüksek olan veri yapıları hakkında fikir verecek şekilde belirtilmiştir.

GDSC verisinde, gen isimleri "*GENE_SYMBOLS*" sütunu altında verilmektedir. "*GENE_title*" sütununda gen ismi açık olarak belirtilmektedir. Hücre hatları "*DATA.*" ön eki bulunan kodlarla sütun isimleri olarak belirtilmiştir. Gen ifade değerleri, hücre hattı sütunları altında bulunmaktadır (Tablo 3.3.).

**Tablo 3.3.** GDSC gen ifade kaynak verisi genel yapısının örneklendirilerek gösterimi.

| GENE_SYMBOLS | GENE_title | DATA.906826 |
|---|---|---|
| TSPAN6 | tetraspanin 6 [Source:HGNC Symbol;Acc:11858] | 7.632023 |
| TNMD | tenomodulin [Source:HGNC Symbol;Acc:17757] | 2.964585 |
| DPM1 | dolichyl-phosphate mannosyltransferase polypep... | 10.37955 |
| SCYL3 | SCY1-like 3 (S. cerevisiae) [Source:HGNC Symbo... | 3.614794 |

CCLE gen ifade verisinde, "*Gene*" sütunu altında gen isimleri belirtilmiştir. Geri kalan sütunlarda hücre hattı isimleri sütun ismi olarak atanmıştır. Gen ifade değerleri hücre hattı sütunları altında bulunmaktadır (Tablo 3.4.).

**Tablo 3.4.** CCLE gen ifade kaynak verisi genel yapısının örneklendirilerek gösterimi.

| Gene | 1321N1 | 143B | 22Rv1 | 23132/87 |
|---|---|---|---|---|
| A1BG | 5.54272 | 5.24589 | 4.75356 | 4.06964 |
| A1BG-AS1 | 4.65816 | 4.56073 | 4.76688 | 4.47164 |
| A1CF | 3.92156 | 3.77696 | 7.44506 | 6.27202 |
| A2M | 4.23675 | 4.50447 | 4.82488 | 3.73186 |
| A2M-AS1 | 3.53547 | 4.37682 | 5.71446 | 4.85001 |

NCI-60 verisinde, "*Probe id*" ve "*Gene name*" sütunlarında gen isimleri belirtilmiştir. "*RefSeq (protein)*" sütunundan sonraki sütunlarda, hücre hattı isimleri önlerinde hangi dokuya ait olduklarını belirten kod (örneğimizde, "BR") ile belirtilmektedir. Gen ifade düzeyleri hücre hattı sütunlarında belirtilmektedir (Tablo 3.5.).

**Tablo 3.5.** NCI-60 gen ifade kaynak verisi genel yapısının örneklendirilerek gösterimi.

| Probe id | Gene name | Entrez gene id | Chromosome | Start | End | Cytoband | RefSeq (mRNA) | RefSeq (protein) | BR: MCF7 |
|---|---|---|---|---|---|---|---|---|---|
| LOC729737 | LOC729737 | 729737 | 1 | 134772 | 140566 | 1p36.33 | NaN | NaN | 9.412 |
| CICP3 | CICP3 | 100132630 | 1 | 656152 | 659631 | 1p36.33 | NaN | NaN | 6.152 |
| LOC101060494 | LOC101060494 | 101060494 | 1 | 661696 | 663628 | NaN | NaN | NaN | 5.435 |
| LINC00115 | LINC00115 | 79854 | 1 | 761585 | 762902 | 1p36.33 | NaN | NaN | 2.296 |
| LINC01128 | LINC01128 | 643837 | 1 | 762970 | 794826 | 1p36.33 | NaN | NaN | 6.737 |

### 3.3. Mutasyon veri tipi

Hücre hattı panellerinin mutasyon veri tipini elde etmek için kullandığı araçlar ve kaynak veri içeriği Tablo 3.6.'da gösterilmiştir.

**Tablo 3.6.** Hücre hattı panellerine ait mutasyon veri tipinin içerikleri ve verinin elde edilmesinde kullanılan araçlar.

| Panel adı | Kaynak | Hücre hattı sayısı | Gen sayısı | Veri elde etmede kullanılan araç / metot |
|---|---|---|---|---|
| GDSC | Orijinal | 1032 | 21972 | Illumina Hiseq 2000 |
| CCLE | Orijinal | 1570 | 19286 | Ekzom yakalama (capture) sekanslama |
| NCI-60 | (29) | 60 | 443 | Ekzom sekanslama |

Aşağıdaki tablolarda; GDSC, CCLE ve NCI-60 mutasyon kaynak veri tablolarının genel yapıları örneklendirilerek sunulmaktadır. Bu örneklendirmelerde gen ve hücre hatlarına ait bölümler, satır ve sütun sayıları yüksek olan veri yapıları hakkında fikir verecek şekilde belirtilmiştir.

GDSC mutasyon verisinde gen isimleri "*gene_symbol*" sütunu altında belirtilmektedir. Protein değişimleri ise "*protein_mutation*" sütununda verilmektedir. Tabloda, hücre hatları ayrı sütunlarda bulunmayıp, "*model_name*" sütununda verilmiştir (Tablo 3.7.).

**Tablo 3.7.** GDSC mutasyon kaynak verisi genel yapısının örneklendirilerek gösterimi.

| gene_id | gene_symbol | model_id | protein_mutation | rna_mutation | cdna_mutation | cancer_driver | model_name |
|---|---|---|---|---|---|---|---|
| SIDG33650 | SAMD11 | SIDM01293 | p.Q413fs*38 | r.1314delC | c.1231delC | FALSE | COLO_099 |
| SIDG42177 | ZNF638 | SIDM00241 | p.T1819fs*4 | r.6077_6078insa | c.5447_5448insA | TRUE | OCUB-M |
| SIDG42177 | ZNF638 | SIDM01154 | p.R1009* | r.3655c>u | c.3025C>T | TRUE | SUP-T1 |
| SIDG42177 | ZNF638 | SIDM00225 | p.Q1555* | r.5293c>u | c.4663C>T | TRUE | IST-MEL1 |
| SIDG42177 | ZNF638 | SIDM01259 | p.? | r.2625+2u>g | c.1995+2T>G | TRUE | GR-ST |

CCLE mutasyon verisinde toplamda 33 sütun bulunmaktadır. Bunlar arasından yöntemimizde kullanılmak üzere oluşturulacak tablo için yalnızca üç sütun dikkate alınmıştır. "*Protein_Change*" sütununda protein değişimleri belirtilmektedir. "*Hugo_symbol*" sütununda ise gen isimleri bulunmaktadır. Hücre hatları ise açık isimleri belirtilmeden "*Broad_ID*" sütununda yer almaktadır (Tablo 3.8.).

**Tablo 3.8.** CCLE mutasyon kaynak verisi genel yapısının örneklendirilerek gösterimi.

| Hugo_Symbol | Entrez_Gene_Id | Chromosome | Start_position | End_position | Protein_Change | Broad_ID |
|---|---|---|---|---|---|---|
| DVL1 | 1855 | 1 | 1277461 | 1277461 | p.E146E | ACH-001270 |
| AL590822.1 | 0 | 1 | 2144416 | 2144416 | p.R202C | ACH-001270 |
| PLCH2 | 9651 | 1 | 2435359 | 2435359 | NaN | ACH-001270 |
| UBE4B | 10277 | 1 | 10177641 | 10177641 | p.E312K | ACH-001270 |

NCI-60 mutasyon verisi, gen temelli olarak mutasyonun varlığı (“1” atanan değeriyle) veya yokluğu (“0” atanan değeriyle) durumlarını belirtecek şekilde oluşturulmuştur. İlk sütunda hücre hatları isimleri, önlerinde ait oldukları dokulara ait kodlar ile belirtilmiştir. Diğer sütun isimlerinde ise gen isimleri bulunmaktadır (Tablo 3.9.).

**Tablo 3.9.** NCI-60 mutasyon kaynak verisi genel yapısının örneklendirilerek gösterimi.

| | **PRDM16** | **CAMTA1** | **MTOR** | **PRDM2** | **CASP9** | **SPEN** |
|---|---|---|---|---|---|---|
| **BR:MCF7** | 0 | 0 | 0 | 0 | 0 | 0 |
| **BR:MDA-MB-231** | 0 | 0 | 0 | 0 | 0 | 0 |
| **BR:HS 578T** | 0 | 0 | 0 | 0 | 0 | 0 |
| **BR:BT-549** | 0 | 0 | 0 | 0 | 0 | 0 |
| **BR:T-47D** | 0 | 0 | 0 | 0 | 0 | 1 |

### 3.4. Metilasyon veri tipi

Hücre hattı panellerinin metilasyon veri tipini elde etmek için kullandığı araçlar ve kaynak veri içeriği Tablo 3.10.‘da gösterilmiştir.

**Tablo 3.10.** Hücre hattı panellerine ait metilasyon veri tipinin içerikleri ve verinin elde edilmesinde kullanılan araçlar.

| **Panel adı** | **Kaynak** | **Hücre hattı sayısı** | **Gen sayısı** | **Veri elde etmede kullanılan araç / metot** |
|---|---|---|---|---|
| GDSC (VİAA) | Orijinal | 790 | 378 | Illumina Human Methylation 450 BeadChip |
| GDSC (ÇAA) | CellMiner | 1080 | 19864 | Illumina Infinium Human Methylation 450 |
| CCLE | CellMiner | 1089 | 19880 | *Reduce representation bisulfite sequencing* (RRBS) |
| NCI-60 | Orijinal | 60 | 17553 | Illumina Infinium Human Methylation 450 |

Aşağıdaki tablolarda; GDSC, CCLE ve NCI-60 metilasyon kaynak veri tablolarının genel yapıları örneklendirilerek sunulmaktadır. Bu örneklendirmelerde gen ve hücre hatlarına ait bölümler, satır ve sütun sayıları yüksek olan veri yapıları hakkında fikir verecek şekilde belirtilmiştir.

Veri içi alan analizinde, GDSC'nin kendi portalında belirtilen orijinal verisi kullanılmıştır. Bu verideki ilk sütunda, daha sonra gen isimlerine çevrilecek olan kromozom üzerindeki koordinasyonlar belirtilmektedir. Sonraki sütunlarda ise hücre hatları, daha sonra açık isimlere çevrilecek olan birer kod numarasıyla belirtilmiştir. Veride, gen temelli olarak metilasyonun olduğu durumlar, var ("1" atanan değeriyle) veya yok ("0" atanan değeriyle) şeklinde gösterilmektedir (Tablo 3.11.).

**Tablo 3.11.** GDSC (orijinal kaynaklı, VİAA için) metilasyon kaynak verisi genel yapısının örneklendirilerek gösterimi.

| Unnamed: 0 | 909703 | 910944 | 906763 | 905952 |
|---|---|---|---|---|
| chr1:10755226-10755521 | 0 | 0 | 0 | 0 |
| chr1:110230238-110230614 | 0 | 1 | 1 | 1 |
| chr1:110880394-110880624 | 0 | 0 | 0 | 0 |
| chr1:111505881-111507007 | 0 | 0 | 0 | 0 |
| chr1:1167001-1168985 | 0 | 0 | 0 | 0 |

Çapraz alan analizinde kullanılmak üzere GDSC'nin CellMiner veri tabanı kaynaklı metilasyon verisi kullanılmıştır. Veride, gen isimleri "*gene*" sütununda belirtilmiştir. Dördüncü sütun olan "*priority*" sütunundan sonra gelen sütunlarda hücre hattı isimleri sütun başlarında kullanılmaktadır. Metilasyon değerleri bu sütunların altında beta değerleri olarak, "0" ve "1" değerleri arasında değer alacak şekilde gösterilmiştir (Tablo 3.12.). Beta değerinin "0" olduğu durum gende metilasyonun olmadığını; "1" olduğu durum ise gende tam bir metilasyonun bulunduğunu göstermektedir.

**Tablo 3.12.** GDSC (CellMiner kaynaklı, ÇAA için) metilasyon kaynak verisi genel yapısının örneklendirilerek gösterimi.

| gene | accession | strand | priority | 201T |
|---|---|---|---|---|
| A1BG | NM_130786 | - | 2 | 0.88807 |
| A1CF | NM_014576;NM_138932;NM_138933 | - | 3 | 0.12089 |
| A2BP1 | NM_001142333;NM_001142333;NM_001142333;NM_0187... | + | 1 | 0.14169 |
| A2LD1 | NM_033110 | - | 1 | 0.77061 |
| A2ML1 | NM_144670 | + | 3 | 0.89617 |

CCLE metilasyon verisinde gen isimleri "*ID*" sütununda belirtilmiştir. Hücre hattı isimleri ise sonraki sütun başlıklarında açık şekilde gösterilmiştir. Veride, gen temelli olarak metilasyon durumları beta değerleri şeklinde hücre hattı sütunlarında sunulmaktadır (Tablo 3.13.).

**Tablo 3.13.** CCLE metilasyon kaynak verisi genel yapısının örneklendirilerek gösterimi.

| ID | 1321N1 | 143B | 22Rv1 | 23132/87 | 253J |
|---|---|---|---|---|---|
| SGIP1 | NaN | NaN | 1 | 0.2829 | NaN |
| AZIN2 | NaN | NaN | 0.20995 | 0.15301 | 0.42214 |
| AGBL4 | NaN | NaN | 0.0763 | 0.65206 | 0.375 |
| NECAP2 | NaN | NaN | 0.37665 | 0.00047 | 0 |
| CLIC4 | NaN | NaN | 0.47216 | 0.44937 | 0.63637 |

NCI-60 metilasyon verisinde gen isimleri "*Gene name*" sütununda bulunmaktadır. "*RefSeq (protein)*" sütunundan sonra gelen sütunların başlıklarında hücre hattı isimleri önlerinde ait oldukları dokuları belirten kodlarla gösterilmiştir. Metilasyon durumları, gen temelli olarak beta değerleri şeklinde hücre hattı sütunlarında sunulmuştur (Tablo 3.14.).

**Tablo 3.14.** NCI-60 metilasyon kaynak verisi genel yapısının örneklendirilerek gösterimi.

| Probe id | Gene name | Entrez gene id | Chromosome | Start | End | Cytoband | RefSeq (mRNA) | RefSeq (protein) | BR: MCF7 |
|---|---|---|---|---|---|---|---|---|---|
| C1orf222 | C1orf222 | 339457 | 1 | -1 | -1 | 1p36.33 | NaN | NaN | 0.851 |
| LOC643837 | LINC01128 | 643837 | 1 | 762970 | 794826 | 1p36.33 | NaN | NaN | 0.049 |
| SAMD11 | SAMD11 | 148398 | 1 | 861120 | 879961 | 1p36.33 | NM_152486.2 | NP_689699.2 | 0.155 |
| NOC2L | NOC2L | 26155 | 1 | 879582 | 894679 | 1p36.33 | NM_015658.3 | NP_056473.2 | 0.528 |
| KLHL17 | KLHL17 | 339451 | 1 | 895966 | 901099 | 1p36.33 | NM_198317.2 | NP_938073.1 | 0.419 |

### 3.5. Kopya Sayısı Değişimi Veri Tipi

Hücre hattı panellerinin KSD veri tipini elde etmek için kullandığı araçlar ve kaynak veri içeriği Tablo 3.15.'te gösterilmiştir.

**Tablo 3.15.** Hücre hattı panellerine ait KSD veri tipinin içerikleri ve verinin elde edilmesinde kullanılan araçlar.

| Panel adı | Kaynak | Hücre hattı sayısı | Gen sayısı | Veri elde etmede kullanılan araç / metot |
|---|---|---|---|---|
| GDSC | Orijinal (PICNIC) | 986 | 24502 | Affymetrix SNP 6.0 |
| CCLE | DepMap (ön işlem yapılmış veri) | 1754 | 25368 | Affymetrix SNP 6.0 |
| NCI-60 | Orijinal (aCGH Agilent 44K) | 60 | 19951 | Agilent Human Genome CGH Microarray 44A; Roche NimbleGen Systems H19 CGH 385K WG Tiling v2.0 array; Affymetrix GeneChip Human Mapping 500 k Array Set; Illumina Human Human1 Mv1_C Beadchip array |

GDSC için KSD verisi olarak, kanserde entegre kopya numaralarını tahmin etme (predicting integral copy numbers in cancer, PICNIC) algoritması kullanılarak oluşturulan mutlak kopya sayıları kullanılmıştır. Veride, "gene_name" sütunu altında gen isimleri

bulunmaktadır. Diğer sütunlardaki başlıklarda ise hücre hattı isimleri açık olarak verilmiştir. KSD değerleri hücre hattı sütunlarında gen temelli olarak belirtilmiştir (Tablo 3.16.).

**Tablo 3.16.** GDSC KSD kaynak verisi genel yapısının örneklendirilerek gösterimi.

| gene_name | M14 | TE-12 | TMK-1 | STS-0421 |
|---|---|---|---|---|
| A1BG | 3.0 | 3.0 | 3.0 | 4.0 |
| A1CF | 3.0 | 3.0 | 3.0 | 4.0 |
| A2M | 3.0 | 3.0 | 2.0 | 4.0 |
| A2ML1 | 3.0 | 3.0 | 2.0 | 4.0 |
| A2ML1-AS1 | 3.0 | 3.0 | 2.0 | 4.0 |

CCLE'ye ait KSD verisi için gen düzeyinde olacak şekilde, kopya sayısı oranına "1" değeri eklenip 2 tabanında logaritması alınarak elde edilen sonuçlar kullanılmıştır. Veride, "gene_name" sütununda gen isimleri belirtilmiştir. Sonraki sütunlarda hücre hattı isimleri parantez içinde ilişkili olan kodlarla beraber verilmiştir. KSD verisi gen temelli olarak hücre hattı sütunlarında sunulmuştur (Tablo 3.17.).

**Tablo 3.17.** CCLE KSD kaynak verisi genel yapısının örneklendirilerek gösterimi.

| gene_name | DDX11L1 (84771) | WASH7P (653635) | MIR6859-1 (102466751) | MIR1302-2 (100302278) |
|---|---|---|---|---|
| ACH-001533 | 1.01613 | 1.01613 | 1.01613 | 1.01613 |
| ACH-000934 | 0.85515 | 0.85515 | 0.85515 | 0.85515 |
| ACH-000653 | 1.06794 | 1.06794 | 1.06794 | 1.06794 |
| ACH-001497 | 1.03922 | 1.03922 | 1.03922 | 1.03922 |
| ACH-000888 | 0.70127 | 0.70127 | 0.70127 | 0.70127 |

NCI-60'ın KSD kaynak verisinin elde edilmesi için, mikrodizi sonuçlarının ön işleme aşamalarından geçirilerek (Agilent Feature Extraction yazılımı, 8.1 versiyonuyla beraber) örneklerdeki DNA ve kontrol DNA ile karşılaştırılmıştır. CellMiner portalından alınan kopya sayısı tahmini (*copy number estimate*) ile etiketlenen KSD verisi çalışmamızda kullanılmıştır. Veride, "Gene name" sütununda gen isimleri belirtilmiştir. "Cytoband" sütunundan sonraki sütunlarda hücre hattı isimleri ait oldukları dokuyu belirten kod ile

beraber verilmiştir. KSD değerleri gen temelli olarak hücre hattı sütunlarında sunulmaktadır (Tablo 3.18.).

**Tablo 3.18.** NCI-60 KSD kaynak verisi genel yapısının örneklendirilerek gösterimi.

| Gene name | Entrez gene id | Chromosome | Start | End | Cytoband | BR: MCF7 |
|---|---|---|---|---|---|---|
| MIG7 | 723788 | 1 | 0 | 0 | 1p22.1 | 1.428 |
| PRO2012 | 55478 | 1 | 0 | 0 | 1q42.13 | 1.454 |
| WASH7P | 653635 | 1 | 14361 | 29370 | 1p36.33 | 2 |
| FAM138A | 645520 | 1 | 34610 | 77690 | 1p36.33 | 2 |
| FAM138F | 641702 | 1 | 34610 | 77690 | 19p13.3 | 2 |

### 3.6. İlaç Tanımlayıcı Veri Tipi

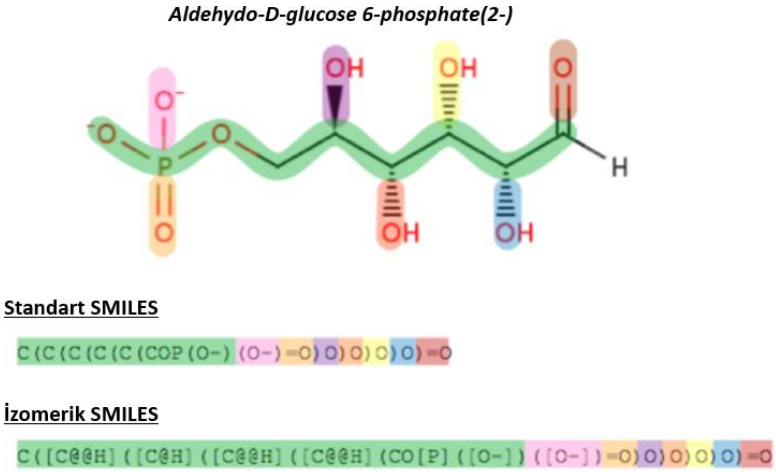
Tasarlanan DeepResponse-RF yönteminde, hücre hattı panellerindeki ilaçların temsili için farklı veri yapılarından yararlanılmıştır. İlaç molekülü yapısının farklı karakterler yardımıyla yeniden ifade edilmesi, makine öğrenmesi algoritmaları tarafından değerlendirilebilme imkanını yaratmıştır.

#### 3.6.1. Basitleştirilmiş Moleküler Girdi Hattı Giriş Sistemi (SMILES)

Yararlanılan veri tiplerinden ilki, SMILES adı verilen formata aittir. SMILES, en yaygın şekilde kullanılan moleküler yapıyı bir karakter dizisi olarak belirtme formatı olarak 1986 yılında David Weininger tarafından geliştirilmiştir (67,68). Kimyasal yapıyı karakter dizisine çevirmekte farklı yöntemler izlenebilmektedir. Bu sebeple, formatın ortaya çıkmasından bu yana standart (canonical) SMILES ve izomerik (isomeric) SMILES gibi yeni format türleri geliştirilmiştir (68).

Bir molekülün farklı şekilde SMILES dizisinin çıkarılması mümkün olması (özellikle kompleks moleküllerde), standart SMILES'ın geliştirilmesine yol açmıştır. Standart SMILES, bahsedilen probleme karşılık olarak her farklı moleküle özel bir SMILES dizisi karşılığı getirmektedir (67,69). Standart SMILES dizisi, bir veri tabanındaki değişik

moleküllerin özgünlüğüne dair bilgiyi aranabilir kılmakta da kullanılmaktadır (67). Şekil 3.1.'de SMILES formatının bir moleküle ait genel görünümü belirtilmiştir.



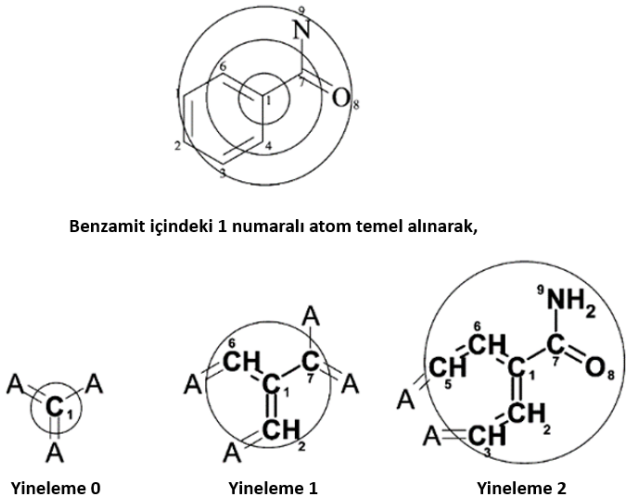

**Şekil 3.1.** Bir moleküle ait SMILES dizisinin gösterimi (70).

Bu tez çalışmasında, veri yapısına erişim ve kullanım kolaylığından dolayı standart SMILES formatı tercih edilmiştir. İlaç moleküllerine karşılık gelen standart SMILES dizileri PubChem (71) ve DrugBank (72) veri tabanlarından elde edilmiştir. Standart SMILES dizisi bulunamayan ilaçlar, kurulan modellerde girdi verisi olarak kullanılmamıştır. SMILES dizisi bulunan ilaçların makine öğrenmesi modeli algoritmaları için anlamlı olması gerekmektedir. Bunun için, SMILES dizileri birer sayısal vektöre dönüştürülerek ilaç parmak izleri çıkartılmıştır.

### 3.6.2. Genişletilmiş Bağlantı Parmak İzleri (ECFP)

İlaç parmak izi (veya moleküler parmak izi, MPİ), iki boyutlu (2B) olarak temsil edilen kimyasal yapıların özellikleri üzerinden elde edilen bir nümerik dizidir (73,74). Bir ilaç parmak izine ait dizi, farklı sayıda öğe (parmak izi tipine göre 16 bit ve 1024 bit arasında değer alabilir) içerecek şekilde düzenlenebilir (74). İlaç parmak izleri, ilk olarak veri tabanlarında arama işlevlerine yardımcı olarak geliştirilse de sonraki dönemlerde benzerlik arama, sanal tarama, sınıflandırma ve kimyasal uzay haritalarının yapılandırılması gibi amaçlar için de kullanılmıştır (75,76).

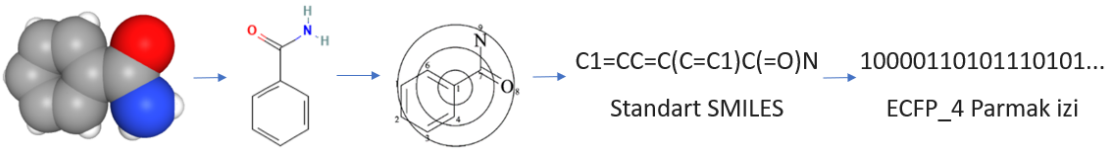
Yakın zamanda geliştirilen Genişletilmiş Bağlantı Parmak İzleri (*Extended Connectivity Fingerprints*, ECFP) adı verilen yöntem, moleküler aktivite ile alakalı yapısal özellikleri elde etme amacıyla tasarlanmıştır. ECFP yöntemi, Morgan algoritması ve Morgan Parmak İzleri yöntemini (77) temel alarak oluşturulmuştur (73,75). Morgan algoritmasındaki yinelemeli işlemde ilk olarak, moleküldeki değişmeyen atom bilgileri ilk atom tanımlayıcısına kodlanır. Devam olarak gelen her adımda, önceki adımlara ait tanımlayıcılar kullanılır. Bu işlemle, tanımlayıcılar orijinal numaralandırmadan bağımsız olarak oluşurlar. Yinelemeler her atom tanımlayıcısının özgün olmasına kadar devam eder (75). ECFP'deki yineleme ise, önceden belirlenmiş sayıya bağlı olarak bitirilir. Her yineleme sonucu oluşan ara tanımlayıcılar atılmadan bir grupta toparlanarak ECFP'yi belirleyen parmak izi oluşturulur (73,75). ECFP için genel bir gösterim ise, kısaltmaya eklenen alt çizgiden sonra gelen sayı iledir. ECFP_4 örneği için, uygulamada 2 yineleme olduğu anlaşılır (75). Şekil 3.2.'de ECFP'deki yineleme adımları örnek olarak bir molekül üzerinden gösterilmiştir.




**Şekil 3.2.** ECFP'de yineleme adımlarının gösterimi (75). Benzamit molekülüne ait 2B grafiği üzerinde dairesel şekilde ilerleyen iki yineleme adımında (ECFP_4, yarıçap = 2) tanımlayıcıların gittikçe daha fazla yapıyı kapsadıkları görülmektedir.

ECFP, yukarıda anlatılan süreçlere bağlı olarak ilaç parmak izi çıkarma için daha az masraflı şekilde hesaplama ve moleküller arası karşılaştırılabilen tanımlayıcıların oluşturulması avantajlarını göstermektedir. Yüksek çıktılı tarama verisi üzerinde de kullanım alanı bulan ECFP, uygulanan Bayesçi temelli yaklaşımla veriden aktif ve inaktif durumların sınıflandırılmasında etkili olduğu raporlanmıştır (75).

ECFP_4, kullanım amacı ilaç aktivitesine yönelik tahminlerde kullanılmaya daha uygun olduğu için DeepResponse-RF'de tercih edilmiştir (75). ECFP üretmek için standart SMILES verisinden yararlanılmıştır. Her ilacın özgün SMILES dizisi karşılığı kullanılarak RDkit kütüphanesinde (78) bulunan metotlar yardımıyla ilaç parmak izleri oluşturulmuştur (Bkz. Ek-2). İlaç parmak izleri, ilgili metot üzerinde yarıçap (radius, yineleme sayısı) değeri 2; dizide bulunan eleman (*bits*) sayısı 1024 olarak belirlenip için ECFP_4 formatında oluşturulmuştur. Şekil 3.3.'te bir moleküle ait SMILES dizisine denk gelen ECFP_4 karşılığına ait örnek belirtilmiştir.

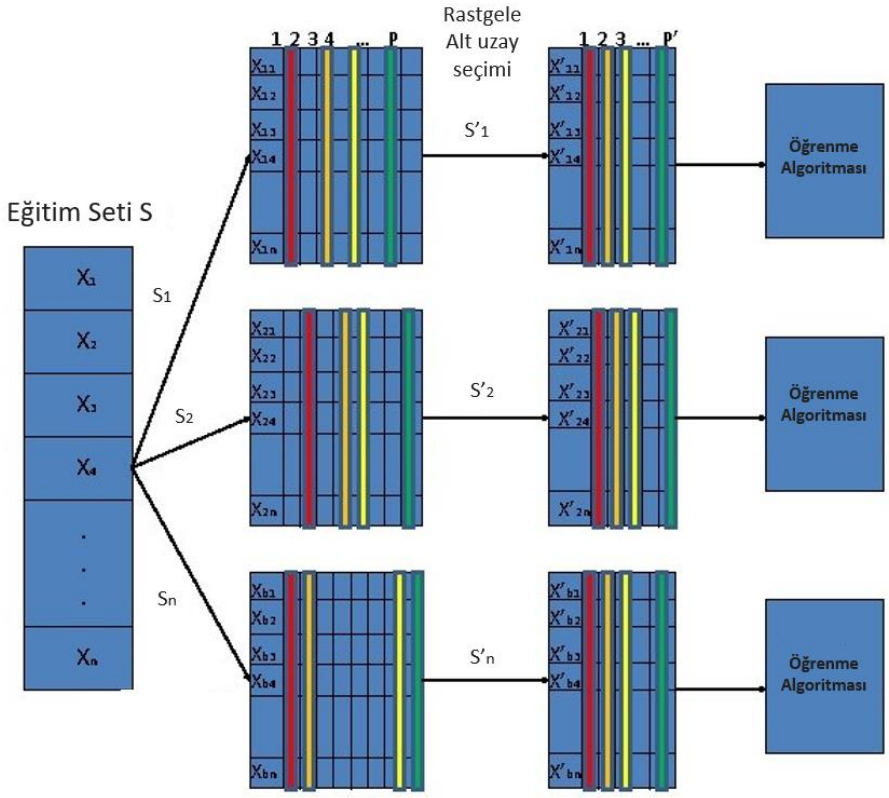



**Şekil 3.3.** Örnek bir molekül (Benzamit) üzerinden SMILES ve ECFP tanımlayıcılarının gösterimi.

### 3.7. Rastgele Orman Algoritması

DeepResponse-RF'nin tasarlanmasında bir makine öğrenmesi yöntemi olan Rastgele Orman (Random Forest, RF) kullanılmıştır. RF, gözetimli öğrenme temelli olarak yakın zamanda geliştirilen yaklaşımların yanında sınıflandırma ve regresyon problemleri için popülerliğini korumaktadır. RF'nin geliştirilmesine yönelik ilk adımlar 1995'te Tin Kam Ho tarafından Rastgele Alt-Uzay Metodu (*Random Subspace Method*, RSM) ile atılmıştır. Topluluk (*ensemble*) metoduna dayanan RSM yönteminde, her biri farklı alt uzaylarda çalışan birkaç sınıflandırıcı (*classifier*) bulunmaktadır (Şekil 3.4.). RSM yöntemi yine Tin tarafından karar ağaçlarına uygulanmıştır (79,80). Sonraki süreçte Leo Breiman

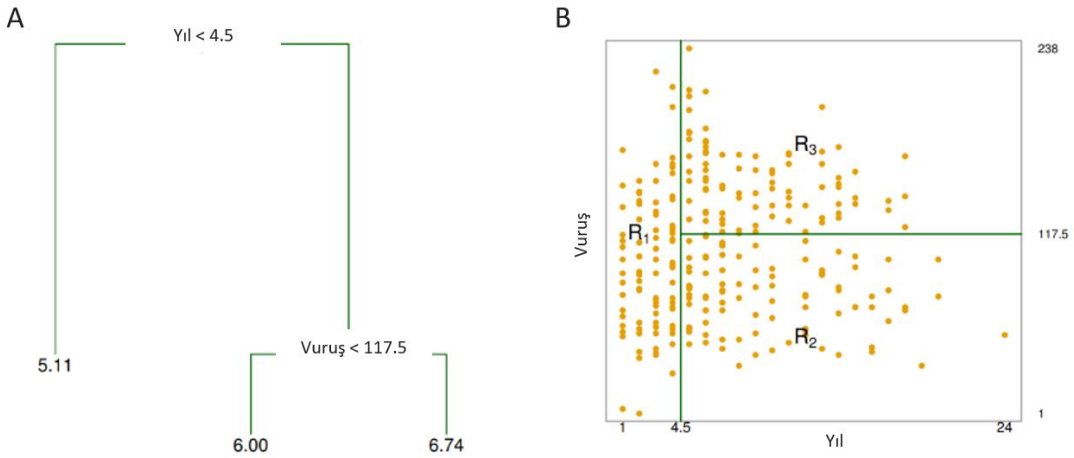
ve Adele Culter, RSM yöntemini torbalama (*bagging*) tekniği ile genişleterek RF yaklaşımını geliştirmişlerdir (79,81).




**Şekil 3.4.** RSM yönteminde uygulanan rastgele alt uzay seçimi aşamaları (82).

Gözetimli bir algoritmaya dayanan karar ağaçları, orijinal veriyi küçük parçalara bölümlemede karar düğümleri kullanır. Algoritma, kök düğümünden (ilk bölümlemenin yapıldığı yer) başlayarak her bölünme aşaması için karar (iç) düğümü oluşturur. İlerlemeler sonucu oluşan şekil ağaç yapısında saçaklı bir form gösterir. Aşağı yönde ilerleyen yapının uç noktalarında yer alan düğümler yaprak (*terminal*) düğümü olarak ifade edilir (Şekil 3.4. (A)) Parametrik olmayan özellikteki RF, bir topluluk metodu olarak birçok karar ağacının bir arada kullanarak iyileştirilmiş sonuçlar sunabilmektedir (83).

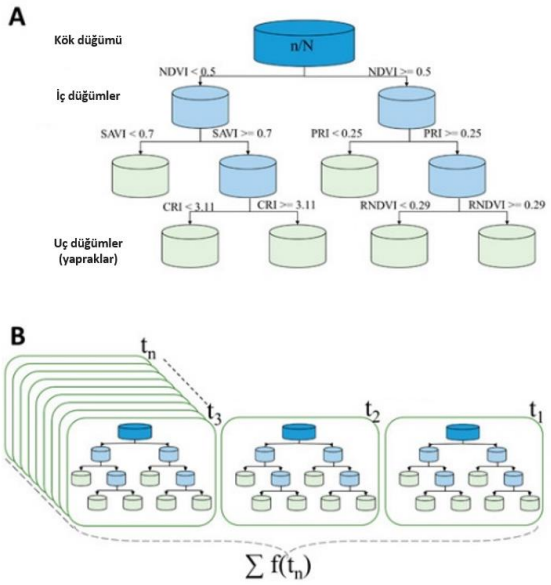
RF'de uygulanan regresyon yönteminde, yeniden örnekleme (*bootstrapping*) tekniğiyle orijinal verinin içinden aynı dağılımı gösteren rastgele örnekler oluşturularak karar ağaçları yapılandırılır.




**Şekil 3.5.** Bir takımdaki sporculara ait veri üzerinden karar ağacı yapısının ortaya çıkarılması (83). Veri, oyuncuların ne kadar maaş alacaklarını belirlemek için ilk ligde geçirdiği yıl ve oyundaki vuruş sayısıyla değerlendirilmiştir. (A)'da verinin 3 katmana ayrılması için uygulanan eşik değerler, kök ve iç düğümler olarak; bölümlemeler sonucu oluşan uç değerler ise yaprak düğümleri görülmektedir. (B)'de ise grafik olarak ayrılan katmanlar veri noktaları ile beraber yansıtılmıştır.

Karar ağacındaki kök düğümü için optimal bölümleyici özellikler arasından tanımlandıktan sonra yeni alt düğümler oluşturulur ve işlem döngüsü yeni düğümler üzerinden (rastgele seçilen düğüm bölümleyici özelliklerle beraber) devam eder. Karar ağacının yapılandırılması yaprak düğüme ulaşıldığında sona erer (Şekil 3.5. (A)) (84). RF'de birden çok karar ağacı kullanıldığı için aşırı öğrenme (*overfitting*) problemi çözülür ve tahmin hatası azaltılmış olur (80,81,83). Sonuçlar, orijinal verinin farklı alt kümeleriyle eğitilen karar ağaçlarının verdiği tahminlerin ortalaması alınarak hesaplanır (Şekil 3.5. (A)).

İlaç yanıtı tahmini problemi için kullanım kolaylığı sunması, stabil olması, hızlı ve doğru şekilde tahmin üretebilmesi avantajlarından dolayı DeepResponse-RF'nin yapılandırılmasında da RF regresyon modeli tercih edilmiştir. Modelin oluşturulma aşamalarında, bir Python programlama dili kütüphanesi olan Scikit-learn (85) kullanılmıştır.




**Şekil 3.6.** RF ile uygulanan regresyon modeli ağacının örnek olarak gösterimi (80). (A) Regresyon için oluşturulan bir karar ağacının gösterimi. (B) Regresyon sonucunun hesaplanmasında, farklı veri alt kümeleriyle eğitilen karar ağaçlarının oluşturduğu tahminlerin ortalaması alınarak sunulur.

### 3.7.1. Modelleme Algoritması ve Hiperparametreler

RF regresyon modelinin algoritması olarak kullanılacak metot için Scikit-learn kütüphanesinde bulunan RandomForestRegressor tercih edilmiştir. Modellemelerde RF içindeki karar ağaçlarının özelliklerini belirlemek için Tablo 3.19.'da tanımları ve atanan değerleri belirtilen max_depth ve n_estimators hiperparametreleri kullanılmıştır.

Veri içi alan analizinde (VİAA), hiperparametrelere ait tüm test değerlerinin (3 max_depth ve 3 n_estimators test değeri için) ikili kombinasyonların olarak 9 farklı senaryo oluşturulmuştur. Buradan alınan performans sonuçlarına göre en iyi performans gösteren ikili değer çapraz alan analizi (ÇAA) ve ablasyon analizinde kullanılmıştır.

**Tablo 3.19.** RF modellemesinde kullanılan hiperparametreler, tanımlar ve test edilen değerler.

| Metot adı | Hiperparametreler | Tanım | Test edilen değerler |
|---|---|---|---|
| sklearn.ensemble.RandomForestRegressor | max_depth | Kök düğüm ve yaprak düğümünün arasındaki en uzun mesafe | 3, 81, 729 |
| | n_estimators | Modelde toplam olarak kurulacak karar ağacı sayısı | 25, 100, 250 |

### 3.7.2. Değerlendirme metrikleri

Modellerin kendi içlerinde değerlendirilmesinde ve diğer yöntemlerin sonuçlarıyla karşılaştırılabilmesi amacıyla 6 farklı metrik kullanılmıştır. Aşağıda bu metriklerin formülleri ve değişkenlerinin tanımları belirtilmiştir.

**Ortalama Mutlak Hata (*Mean Absolute Error*, MAE)**

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{3.1}$$

$y_i$ Tahmin değeri
$x_i$ Gerçek değer
$n$ Toplam veri noktası sayısı

**Ortalama Kare Hata (*Mean Squared Error*, MSE)**

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{3.2}$$

$n$ Toplam veri noktası sayısı
$Y_i$ Gerçek değer
$\hat{Y}_i$ Tahmin değeri

**Kök Ortalama Kare Hatası (*Root Mean Squared Error*, RMSE)**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Yi - \hat{Y}i)^2} \quad (3.3)$$

$n$ Toplam veri noktası sayısı
$Y_i$ Gerçek değer
$\hat{Y}_i$ Tahmin değeri

**Spearman'ın Sıra Korelasyon Katsayısı (*Spearman's Rank Correlation Coefficient*, SCC)**

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (3.4)$$

ρ Spearman'ın sıra korelasyon katsayısı
$d_i$ Karşılıklı veri elemanlarına ait sıra değerlerinin arasındaki fark
$n$ Toplam veri noktası sayısı

**Pearson Korelasyon Katsayısı (*Pearson Correlation Coefficient*, PCC)**

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (3.5)$$

$r$ Korelasyon katsayısı
$x_i$ Gerçek değerler içindeki bir değer
$\bar{x}$ Gerçek değerlerin ortalaması
$y_i$ Tahmini değerler içindeki bir değer
$\bar{y}$ Tahmin değerlerinin ortalaması

**Kararlılık Katsayısı (*Coefficient of Determination*, $R^2$)**

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y}_i)^2} \quad (3.6)$$

SSRES Artık değer karelerinin toplamı
SSTOT Ortalamaya Uzaklığın Kareler Toplamı
$y_i$ Gerçek değer
$\hat{y}$ Tahmin değeri
$\overline{y_i}$ Ortalama değer

### 3.7.3. Uygulama Detayları ve Kullanılan Araçlar

Çalışmaya ait kodlar Python programlama dili (sürüm 3.8.5) ile yazılmıştır. Girdi verilerinin hazırlanmasında Pandas (86), NumPy (87), SciPy (88), Scikit-learn (85), RDKit (78), PubChemPy (89) kütüphanelerinden ve PUG-REST (90) programlı erişim yolundan yararlanılmıştır. Verilerin görselleştirilmesinde Matplotlib (91) ve Seaborn (92) kütüphaneleri kullanılmıştır. Modelin programlanmasında Scikit-learn kütüphanesinde yer alan RandomForestRegressor metodu kullanılmıştır.

Veri tiplerinin ön işleme aşamalarında kişisel bilgisayardan yararlanılmıştır. Model eğitimi, doğrulama, optimizasyonu ve kapsamlı tahmin üretimi aşamaları için GPU'lu sunucu kullanılmıştır. GPU'lu sunucu özellikleri; HP Z8 G4 iş istasyonu (2 x HP Intel Xeon Gold 5215 2.50 GHz CPU, 20 çekirdek, 40 şerit), HP 64 GB (4 x 16GB) DDR4 2933 MHz bellek, GPU: 2 x NVIDIA GeForce RTX 2080 Ti 11 GB şeklindedir.

## 3.8. Uygulanan Analizler Ve Veri Ön İşleme Aşamaları

### 3.8.1. Hücre Hattı Özellik Veri Tiplerinin Oluşturulması

Farklı hücre hattı panellerinden alınan kaynak veri dosyaları aynı formatta düzenlenip veri tiplerine ait tablolar oluşturulmuştur. Tüm hücre hattı özellik veri tipi tabloları, ilk sütunda gen isimleri; sonraki sütunlarda ise hücre hatlarının ilgili gen için atanan değeri bulunmaktadır. Hem indeks olarak bulunan gen ismi sütunu hem sütun başlarındaki hücre hattı isimleri alfabetik olarak sıralandıktan sonra kaydedilmiştir.

Ek olarak, çapraz alan analizinde, tüm veri tiplerinde bulunan gen, hücre hattı isimlerinin karşılaştırılabilir olması için bu isimler ortak formatta düzenlenmiştir. İsimler, harf ve sayı harici karakterler silindikten sonra küçük harfle tekrar yazdırılıp kaydedilmiştir.

Kaynak veri yapıları üzerinden edilip belirli bir formatta düzenlenen omik veri tipleri üzerinden genel yapı hakkında fikir vermesi açısından örneklemeler yapılmıştır.

Tüm omik veri tipleri kendi aralarında benzer yapılara sahip olduğundan örnekleme gösterimleri yalnızca GDSC verisi üzerinden sunulmuştur. Tablo 3.20. – 3.24.'te sırasıyla GDSC verisine ait gen ifade, mutasyon, metilasyon (iki değerlikli), metilasyon (beta değerli), KSD veri tiplerine ait örneklemeler gösterilmektedir.

**Tablo 3.20.** GDSC gen ifade ön işlemli verisi genel yapısının örneklendirilerek gösterimi.

| gene_name | 201T | 22RV1 | 23132-87 | 42-MG-BA |
|---|---|---|---|---|
| A1BG | 3.162073 | 3.530584 | 3.361207 | 6.0023 |
| A1CF | 2.919189 | 6.364383 | 6.292324 | 3.17314 |
| A2M | 3.379281 | 5.333116 | 3.473164 | 3.23721 |
| A2ML1 | 2.700304 | 2.907726 | 2.815961 | 2.79058 |
| A3GALT2P | 2.674981 | 2.815383 | 2.913369 | 2.6262 |

**Tablo 3.21.** GDSC mutasyon ön işlemli verisi genel yapısının örneklendirilerek gösterimi.

| gene_name | COLO_099 | HGC-27 | NCI-H2172 | KURAMOCHI |
|---|---|---|---|---|
| SAMD11 | 1 | 0 | 0 | 0 |
| AGRN | 1 | 0 | 0 | 0 |
| CFAP74 | 0 | 1 | 0 | 1 |
| RPL22 | 1 | 0 | 0 | 0 |
| VPS13B | 0 | 1 | 0 | 0 |

**Tablo 3.22.** GDSC metilasyon (iki değerlikli) ön işlemli verisi genel yapısının örneklendirilerek gösterimi.

| gene_name | MY-M12 | RS4-11 | ALL-PO | BE-13 | CCRF-CEM |
|---|---|---|---|---|---|
| ABCB1 | 0 | 0 | 0 | 0 | 0 |
| ACCS | 0 | 0 | 0 | 0 | 0 |
| ACP5 | 0 | 0 | 0 | 0 | 0 |
| ADAM11 | 0 | 0 | 0 | 0 | 0 |
| ADAM19 | 0 | 0 | 0 | 0 | 0 |

**Tablo 3.23.** GDSC metilasyon (beta değerli) ön işlemli verisi genel yapısının örneklendirilerek gösterimi.

| gene_name | 201T | 22RV1 | 23132-87 | 42-MG-BA |
|---|---|---|---|---|
| A1BG | 0.88807 | 0.88047 | 0.05773 | 0.9287 |
| A1CF | 0.12089 | 0.32279 | 0.09157 | 0.11224 |
| A2BP1 | 0.14169 | 0.15159 | 0.35441 | 0.2603 |
| A2LD1 | 0.77061 | 0.26594 | 0.53436 | 0.78356 |
| A2ML1 | 0.89617 | 0.86326 | 0.31267 | 0.89576 |

**Tablo 3.24.** GDSC KSD ön işlemli verisi genel yapısının örneklendirilerek gösterimi.

| gene_name | M14 | TE-12 | TMK-1 | STS-0421 |
|---|---|---|---|---|
| A1BG | 3.0 | 3.0 | 3.0 | 4.0 |
| A1CF | 3.0 | 3.0 | 3.0 | 4.0 |
| A2M | 3.0 | 3.0 | 2.0 | 4.0 |
| A2ML1 | 3.0 | 3.0 | 2.0 | 4.0 |
| A2ML1-AS1 | 3.0 | 3.0 | 2.0 | 4.0 |

### 3.8.2. İlaç Yanıtı Verisinin Düzenlenmesi

İlaç yanıtı verisi için genel olarak tablolar 3 sütundan oluşturulmuştur. Hücre hattı isimleri, İlaç isimleri, pIC50 (NCI-60'ta pIC50 yerine pGI50 sütunu bulunmaktadır). Veri içi alan analizinde isim sütunlarında bir değişikliğe gidilmeye gerek görülmemiştir. Ancak, çapraz alan analizinde isimlerin karşılaştırılabilir olması açısından isimler, harf ve sayı harici karakterler silindikten sonra küçük harfle tekrar yazdırılıp kaydedilmiştir. Buna bağlı olarak, çapraz alan analizinde asıl isim sütunlarının yanı sıra düzenlenmiş olanlara ait yeni sütunlar eklenmiştir.

Hücre hattı panellerinin genel olarak kullandığı IC50 cinsindeki ilaç yanıtı değerleri lineer skalada değişmektedir. Bundan dolayı, yanıt değerleri arasında 0 veya negatif değerler de görülebilmektedir. IC50 değerlerinin negatif logaritmaları alınarak hesaplanan pIC50 ise logaritmik skalada ilaç etkisinin daha doğru karşılaştırılabilir olmasını sağlamaktadır. Bu tez çalışması kapsamında da ilaç yanıtı değerleri için pIC50 değerlerinin kullanılması uygun görülmüştür.

pIC50 değerlerinin hesaplanmasında izlenen yöntemde, ilk olarak mikromolar cinsinde olan IC50 değerleri $10^6$ değerine bölünür. Sonrasında, bölüm değerinin 10 tabanında negatif logaritma değeri hesaplanarak pIC50 değeri bulunur.

Aşağıda, panellere ait ilaç yanıtı veri yapılarının oluşturulması için kullanılan kaynaklar ve onların genel içerikleri belirtilmiştir (Tablo 3.25.).

**Tablo 3.25.** İlaç yanıtı verisi için kaynak olan dosyaların özellikleri.

| Panel Adı | Kaynak | Kullanılan Yöntem | Birim |
|---|---|---|---|
| GDSC | Orijinal | Syto60 ve Resazurin (GDSC1), CellTiter Glo (GDSC2) | IC50 (µM) |
| CCLE | Orijinal | CellTiter Glo | IC50 (µM) |
| NCI-60 | Orijinal (rcellminer) | CellTiter Glo | GI50 (µM) |

GDSC ilaç yanıtı verisini oluşturmak için kullanılan GDSC1 ve GDSC2 versiyonlarının ikisinde de aynı sütunlar kullanılmıştır. Veride kullanılan sütun sayısı 19 olduğundan, örnek olarak belirli sütunlar yalnızca GDSC1 üzerinden örneklenmiştir. Tablo 3.26.'da hücre hattı ve ilaç isimleri sırasıyla "*CELL_LINE_NAME*" ve "DRUG_NAME" sütunlarında belirtilmektedir. Panelde incelenen her çiftin deneysel veri noktası (IC50), doğal logaritması alınmış şekilde "*LN_IC50*" sütununda verilmiştir.

**Tablo 3.26.** GDSC ilaç yanıtı kaynak verisi genel yapısının örneklendirilerek gösterimi.

| DATASET | NLME_ RESULT_ ID | NLME_ CURVE_ID | COSMIC_ ID | CELL_LINE _NAME | DRUG_ ID | DRUG_ NAME | LN_IC50 |
|---|---|---|---|---|---|---|---|
| GDSC1 | 281 | 12974350 | 683665 | MC-CAR | 1 | Erlotinib | 2.39569 |
| GDSC1 | 281 | 12975300 | 684055 | ES3 | 1 | Erlotinib | 3.14092 |

**Tablo 3.26. (Devam)** GDSC ilaç yanıtı kaynak verisi genel yapısının örneklendirilerek gösterimi.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| GDSC1 | 281 | 12975647 | 684057 | ES5 | 1 | Erlotinib | 3.96876 |
| GDSC1 | 281 | 12975980 | 684059 | ES7 | 1 | Erlotinib | 2.692768 |
| GDSC1 | 281 | 12976330 | 684062 | EW-11 | 1 | Erlotinib | 2.478678 |

CCLE ilaç yanıtı kaynak verisinde 18 sütun bulunmaktadır. Verinin örneklenmesi için çalışmamız kapsamında göz önüne aldığımız bazı sütunlar kullanılmıştır. Tablo 3.27.'de, hücre hattı ve ilaç isimleri sırasıyla "*Primary Cell Line Name*", "*Compound*" sütunlarında verilmiştir. Panelde incelenen her çift için ölçülen deneysel ilaç yanıtı değeri "IC50 (μM)" sütununda gösterilmektedir.

**Tablo 3.27.** CCLE ilaç yanıtı kaynak verisi genel yapısının örneklendirilerek gösterimi.

| **CCLE Cell Line Name** | **Primary Cell Line Name** | **Compound** | **IC50 (μM)** |
|---|---|---|---|
| 1321N1_CENTRAL_NERVOUS_SYSTEM | 1321N1 | AEW541 | 8 |
| 22RV1_PROSTATE | 22Rv1 | AEW541 | 2.32992 |
| 42MGBA_CENTRAL_NERVOUS_SYSTEM | 42-MG-BA | AEW541 | 2.68213 |
| 5637_URINARY_TRACT | 5637 | AEW541 | 5.00231 |
| 639V_URINARY_TRACT | 639-V | AEW541 | 1.73618 |

NCI-60 ilaç yanıtı verisindeki ilk sütunda ilaçların kimlikleri olan NSC (Ulusal Hizmet Merkezi, *National Service Center*) numarası bulunmaktadır. Diğer sütunlarda ise hücre hattı isimleri, ait oldukları dokuların kodlarını gösteren ön eklerle beraber

belirtilmiştir. İlaç ve hücre hattı çiftlerinin deneysel veri noktaları (GI50) hücre hattı sütunlarında ilaç temelli olarak sunulmuştur (Tablo 3.28.).

**Tablo 3.28.** NCI-60 ilaç yanıtı kaynak verisi genel yapısının örneklendirilerek gösterimi.

| NSC | BR:MCF7 | BR:MDA-MB-231 | BR:HS 578T | BR:BT-549 | BR:T-47D |
|---|---|---|---|---|---|
| 26980 | 6.84192 | 5.16962 | 5.4593 | 5.67097 | 5.74616 |
| 134727 | 4.79582 | 3.63216 | 3.70363 | 4.26387 | 4.22001 |
| 755880 | 7.092 | 5.04803 | 5.1995 | 5.88596 | 5.7765 |
| 1895 | 2.37095 | 2 | NaN | NaN | 2 |
| 56410 | 6.31182 | 4.39248 | 4.79172 | 4.96614 | 4.74506 |

Aşağıda sunulan tablolarda GDSC, CCLE ve NCI-60 kaynak tabloları kullanılarak üretilen ön işlemli ilaç yanıtı veri tipleri bulunmaktadır (Tablo 3.29. – 3.31.).

**Tablo 3.29.** GDSC ilaç yanıtı ön işlemli verisi genel yapısının örneklendirilerek gösterimi.

| DRUG_NAME | CELL_LINE_NAME | pIC50 |
|---|---|---|
| (5Z)-7-Oxozeaenol | 22RV1 | 4.925703 |
| (5Z)-7-Oxozeaenol | 23132-87 | 5.434868 |
| (5Z)-7-Oxozeaenol | 42-MG-BA | 5.744347 |
| (5Z)-7-Oxozeaenol | 451Lu | 7.441469 |
| (5Z)-7-Oxozeaenol | 5637 | 5.293581 |

**Tablo 3.30.** CCLE ilaç yanıtı ön işlemli verisi genel yapısının örneklendirilerek gösterimi.

| DRUG_NAME | CELL_LINE_NAME | pIC50 |
|---|---|---|
| AEW541 | 1321N1 | 5.09691 |
| AEW541 | 22Rv1 | 5.632658 |
| AEW541 | 42-MG-BA | 5.57152 |
| AEW541 | 5637 | 5.300829 |
| AEW541 | 639-V | 5.760405 |

**Tablo 3.31.** NCI-60 ilaç yanıtı ön işlemli verisi genel yapısının örneklendirilerek gösterimi.

| DRUG_NAME | NSC_ID | CELL_LINE_NAME | pGI50 |
|---|---|---|---|
| Methotrexate | 740 | BT-549 | 7.54792 |
| Methotrexate | 740 | HS 578T | 4.7776 |
| Methotrexate | 740 | MCF7 | 3.85963 |
| Methotrexate | 740 | MDA-MB-231 | 5.0659 |
| Methotrexate | 740 | T-47D | 4.7453 |

**İlaç Tanımlayıcı Veri Tipinin Oluşturulması**

İlaç yanıtı veri tiplerinde yer alan ilaçlara karşılık gelen standart SMILES dizilerinin bulunması için PubChem PUG-REST programlı erişimi, PubChemPy kütüphanesi ve DrugBank (sürüm 5.1.6) ilaç bilgi veri setinden yararlanılmıştır.

Standart SMILES dizilerinin ilaç parmak izlerine çevrilmesi için RDKit kütüphanesi kullanılmıştır. Bazı ilaçlara karşılık gelen SMILES dizileri olmadığı için ilaç parmak izleri çıkartılmamıştır. İlaç tanımlayıcı verisi olmayan bu moleküllere ait ilaç yanıtı veri noktaları değerlendirme dışında tutulmuştur.

İlaç parmak izlerinin oluşturulmasında RDKit metodu olan MolFromSmiles ve GetMorganFingerprintAsBitVect kullanılmıştır. İlk olarak, MolFromSmiles ile bir SMILES dizisi için obje oluşturulur. Ardından, GetMorganFingerprintAsBitVect metodu yardımıyla bu obje üzerinden ECFP_4 formatında ve 1024 bit uzunluğundaki ilaç parmak izi elde edilir. Tablo 3.32.'de kullanılan parametre ve atanan değer belirtilmiştir.

**Tablo 3.32.** İlaç parmak izlerinin çıkarılması kullanılan metotlar ve parametre tercihleri.

| Metot adı | Metot Amacı | Parametre adı | Tanım | Atanan değer |
|---|---|---|---|---|
| MolFromSmiles | SMILES dizisi için bir obje oluşturulması | - | - | - |
| AllChem.GetMorganFingerprintAsBitVect | SMILES objesi kullanılarak ECFP formunda ilaç parmak izi üretilmesi | radius | Yarıçap, parmak izinin kaç adımda oluşturulacağı değer | 2 |
| | | nBits | Parmak izinin bulunduracağı eleman sayısı (bit) | 1024 |

Modelleme sırasında 1024 bit uzunluğundaki ilaç parmak izleri ayrı ayrı şekilde 1024 farklı sütun değeri olarak hücre hattı özellik veri tipleriyle oluşturulan vektörün devamına eklenmiştir.

### 3.8.3. Hücre Hattı Özellik Vektörlerinde Kullanılacak Gen Sayısının Azaltılması

Tahmin performansının artırımı ve model eğitimi süresinin minimize edilmesi amacıyla hücre hattı özellik vektörlerine eklenen gen sayılarının azaltılması için bazı düzenlemeler yapılmıştır. Hücrelerdeki gen ifade profilini daha iyi temsil ettiği bildirilen 978 genin oluşturduğu L1000 listesinden (31) bu amaçla yararlanılmıştır.

Öncelikle, hücre hattı özellik veri tiplerinin genişletilmiş tabloları (27794 gen, 1126 hücre hattı içeren tablolar) üzerinden yapılan doluluk oranı (gen ifade ve KSD verisi için) ve 1 sayısı bulundurma oranı (mutasyon ve metilasyon verisi için) analizi sonrası belirli eşik değerleri kullanılıp gen sayıları azaltılmıştır. Uygulanan eşik değerleri ve kalan gen sayıları sırasıyla şu şekildedir; gen ifade (% 90, 16468), mutasyon (% 5, 2458), metilasyon (% 1, 148), KSD (% 87, 14648). Elde edilen bu gen sayılarının bir vektörde birleştirilip kurulacak modelde girdi verisi olarak kullanımı, harcanacak zaman ve modelin tahmin performansı açısından uygulanabilir olmadığı sonucuna varılmıştır. Bu nedenle gen sayılarının azaltımı için bir filtrelemeye daha ihtiyaç duyulmuştur.

Yapılan ek işlemde, yukarıda belirtilen veri tiplerine ait gen listelerinden L1000 genleri filtrelenmiştir. Kullanılan ana liste ve filtreleme sonucu oluşan gen liste uzunlukları sırasıyla şu şekildedir; gen ifade (16468, 896), mutasyon (2458, 107), metilasyon (148, 11), KSD (16468, 896). Bu sonuçlarla, bir hücre hattı özellik vektörünün uzunluğu 1910 olmuştur.

Ana gen listesi (35542 vektör uzunluklu) ve yeni gen listesiyle (1910 vektör uzunluklu) oluşturulan tabloların karşılaştırılabilmesi için sindirim sistemi dokusuna ait hücre hatları seçilip 5 ayrı ilaç yanıtı tahmini analizi yapılmıştır (Tablo 3.34.). Karşılaştırmalarda, GridSearchCV, cross_val_score, RandomForestRegressor (RF tahmin edicisi) metotları 5 kat çapraz doğrulama yapılabilmesi için, Tablo 3.33.'de belirtilen hiperparametreler ve onlara atanan değerlerle kullanılmıştır. Performanslar, analiz süresi ve MAE skorlama metriği üzerinden karşılaştırılmıştır.

**Tablo 3.33.** Gen azaltma aşamasında model karşılaştırmaları için kullanılan hiperparametreler, onlara ait tanımlar ve atanan değerler.

| Hiperparametre adı | Tanım | Atanan değer |
|---|---|---|
| bootstrap | Bootstrap örnekleme yöntemiyle elde edilen verinin karar ağaçlarının oluşturulmasında kullanımı | *True* |
| min_samples_leaf | Bir yaprak düğümünde olması gereken minimum örnek sayısı. Herhangi bir derinlikte bölünmüş bir nokta, yalnızca sol ve sağ dalların her birinde en az min_samples_leaf eğitim örnekleri bırakırsa dikkate alınır. | 1 |
| min_samples_split | Bir iç düğümün bölünmesi için gereken minimum örnek sayısı | 2 |
| max_ features | En iyi bölünmeyi ararken dikkate alınması gereken özellik sayısı | *Auto* |
| oob_score | Genelleme puanını tahmin etmek için out-of-bag örneklerin kullanılıp kullanılmayacağı durumu | *True* |

**Tablo 3.34.** Gen azaltma aşamasında model karşılaştırmaları için kullanılan veri tipleri ve hiperparametre seçimleri.

| Kullanılan veri | Hiperparametre Seçimleri |
|---|---|
| Sindirim sistemi (35542 vektör uzunluklu) | "bootstrap" : [True], "max_depth" : [5], "min_samples_leaf" : [1], "min_samples_split" : [2], "n_estimators" : [100], "max_features" : ["auto"] |
| Sindirim sistemi (35542 vektör uzunluklu) | "bootstrap" : [True], "max_depth" : [5], "min_samples_leaf" : [1], "min_samples_split" : [2], "n_estimators" : [25], "max_features" : ["auto"], "oob_score" : [True] |
| Sindirim sistemi (35542 vektör uzunluklu) | "bootstrap" : [True], "max_depth" : [5], "min_samples_leaf" : [1], "min_samples_split" : [2], "n_estimators" : [100], "max_features" : ["auto"], "oob_score" : [True] |
| Sindirim sistemi (1910 vektör uzunluklu) | "bootstrap" : [True], "max_depth" : [5], "min_samples_leaf" : [1], "min_samples_split" : [2], "n_estimators" : [100], "max_features" : ["auto"], "oob_score" : [True] |
| Sindirim sistemi (1910 vektör uzunluklu) | "bootstrap" : [True], "max_depth" : [5], "min_samples_leaf" : [1], "min_samples_split" : [2], "n_estimators" : [25], "max_features" : ["auto"], "oob_score" : [True] |

### 3.9. T-SNE Veri Görselleştirmesi

Oluşturulan hücre hattı özellik veri tipleri, içerdikleri çok sayıda gen ve hücre hattı değişkenleri nedeniyle çok boyutlu veri kümeleridirler. Veri tiplerinin görselleştirilmesi, barındırılan özelliklerin tanınabilmesi açısından önemlidir. Buradan hareketle, bir lineer olmayan boyut azaltma yöntemi olan t-dağıtılmış Stokastik Komşu Gömme (*t-distributed Stochastic Neighbor Embedding*, t-SNE) (93), yüksek sayıda özellik içeren veri tiplerinin iki boyutta ifade edilebilmesi için kullanılmıştır.

tSNE grafiklerinin oluşturulması için izlenen yöntemde her bir omik veri tipi için ve bunların bir arada olduğu veri ayrı ayrı değerlendirilerek beşer adet grafik oluşturulmuştur. Düzenleme için seçilen veri yapıları, önceden oluşturulan kaynak veri tipleri, 35542 ve 3747 hücre hattı özellik vektörü uzunluğuna sahip tablolardır. Ayrıca, hücre hatlarının doku bazında gruplanabilmesi için, 986 hücre hattına atfedilen doku isimlerinin bulunduğu dosya kullanılmıştır. Düzenleme aşamalarında, Sklearn, NumPy, Pandas, Matplotlib, Seaborn, Distinctipy (94) gibi Python kütüphanelerinden yararlanılmıştır.

Gen ifade verisine ait grafiği çıkarmak için uygulanan adımlar diğer omik veri tiplerinde de aynı olduğu için diğerlerinde sadece farklı değerlendirilen noktalar belirtilmiştir.

Her bir omik veri tipi için ayrı ayrı ilerletilen süreçte, veriye (tekil omik veya birleştirilmiş tip) ait sütunları belirten sütunlar ve doku numaraları sütunu ayrı değişkenlere atanmıştır. TSNE metodu n_components (hedef boyutların sayısı) parametre ve atanan “2” değeriyle kullanılarak bir obje oluşturulmuştur. Ardından, vektör verisi obje yardımıyla gömülü uzaya uydurulup dönüştürülen çıktısı alınmıştır. Elde edilen çıktı, boş bir tabloya üç sütunlu halde aktarılmıştır. İlk sütundaki dokulara ait sayılar ters çevrilmiş sözlük yardımıyla düzenlenip doku isimleriyle değiştirilmiştir. Grafik üzerinde renklerin ayırt edilebilmesi amacıyla Distinctipy kütüphanesi yardımıyla 13 adet

birbirinden farklı renk seçilmiştir. Seaborn kütüphanesinden relplot metodu kullanılarak son elde edilen tablo ve renk kartelası kullanılıp saçılım grafik türü seçilerek tSNE grafiği elde edilip kaydedilmiştir. Geniş ve ayrıntılı anlatım EK-2'de yapılmıştır.

### 3.10. Vektör Matrisinin Oluşturulması

Modelde girdi verisi olarak kullanılacak olan vektör matrisleri, veri tiplerinin genişletilmiş formattaki tabloları kullanılarak oluşturulmuştur. Vektör matrisinin her bir satırında farklı bir hücre hattına ait özellik veri tiplerinin yatay olarak birleştirilmiş hali bulunur. Hücre hattı özellik veri tiplerinin birleştirilmesinde, gen ifade, mutasyon, metilasyon, kopya sayısı değişimi sırası benimsenmiştir.

Veri içi alan analizi için hazırlanan vektör matrisinin oluşturulmasında veri tiplerinin genişletilmiş formatta olanları kullanılmıştır. Veri tablolarında yer alan boşluk değerlerinin doldurulması için gen bazında hesaplamalar gen ifade verisi için ortalama; KSD için medyan değeri kullanılmıştır. Mutasyon ve metilasyon veri tipleri içindeki boşluklar ise 0 (yok) değeriyle doldurulmuştur.

Veri tiplerindeki tüm genlerin kullanılması modelleme yaklaşımı açısından uygun olmadığından, vektör uzunluğunun kısaltılması için düzenlemeler yapılmıştır. Veri tipi başına düşen gen miktarını azaltmak için gen ifade ve KSD tablolarındaki satırların doluluk oranları; mutasyon ve metilasyon tablolarındaki satırlarda 1 değerinin yüzde olarak ne kadar bulunduğu hesaplanmıştır. Her gen için hesaplanan yüzde değerleri üzerine eşik değerleri (% 90 gen ifade, % 5 mutasyon, % 1 metilasyon, % 87 KSD için) uygulanarak eşiğin altında değere sahip olan genler silinmiştir. Ardından, her listenin L1000 gen listesi ile ortaklıkları bulunduktan sonra, veri tipi başına düşen gen sayısı şu şekilde olmuştur, 923 (gen ifade), 110 (mutasyon), 12 (metilasyon), 944 (KSD). Bir hücre hattına ait bir vektör uzunluğu ise bu gen listelerinin uzunluğu olan 1989 olarak hesaplanmıştır.

Çapraz alan analizinde farklı panellere ait veri yapıları değerlendirileceği için vektör boyutu belirli bir sayıda sabitlenmiştir. Bunu sağlamak için, panellerdeki aynı tür

veri tiplerinin önceden düzenlenen gen isim listeleri kullanılarak her türün ortak gen isim listeleri oluşturulmuştur. Devamında, tüm veri tipi listelerinin L1000 gen listesiyle ortaklıkları bulunarak listeler son halini almıştır. Sonuç olarak elde edilen gen isim listelerinin eleman sayıları ve listelerin birleştirilmesiyle oluşan bir vektörün uzunluk değeri sırasıyla şöyledir; 897 (gen ifade), 963 (mutasyon), 955 (metilasyon), 932 (KSD), 3747.

Vektör matrisleri oluşturulurken veri içi alan analizinde izlenen yöntem akışı aynı şekilde burada da uygulanmıştır. GDSC metilasyon verisi farklı kaynaktan alındığı ve beta değerlerini içerdiği için tablo boşlukları 0 yerine beta değerlerinin ortalamasıyla doldurulmuştur. Yüzde eşik değerleri, , % 87 gen ifade, % 80 metilasyon, % 75 KSD olarak belirlenmiştir (mutasyon verisi için eşik değeri uygulanmamıştır). GDSC, CCLE, NCI-60 panellerinin vektör boyutları aynı genler kullanıldığı için eşit uzunlukta olmuştur.

Ablasyon analizi için kullanılan birleştirilmiş omikler veri tiplerine ait vektör matrisi çapraz alan analizinde kullanılan GDSC matrisiyle aynıdır. Ancak, veri tiplerinin ayrı ayrı değerlendirildiği senaryolarda vektör uzunluğu yukarıda belirtilen ortaklık listelerinin eleman sayılarına eşittir.

### 3.11. Analiz Tipleri

#### 3.11.1. Ablasyon Analizi

GDSC paneli için, her veri tipinin ve bunların birleştirilmiş formunun ilaç yanıtı tahminindeki etkisini ayrı senaryolar üzerinden görebilmek için ablasyon analizi yapılmıştır. Tablo 3.35.‘te ablasyon analizi için kullanılan tablolar ve genel özellikleri belirtilmiştir. Şekil 3.7.’de ablasyon analizinde uygulanan işlemlerin genelleyici görselleştirilmesi yapılmıştır.

Aşağıda, omik veri tipleri ve birleştirilmiş omikler verisi için uygulanan analiz aşamaları belirtilmiştir. Tüm analizlerde uygulanan yöntemler izlenip, sadece kullanılan

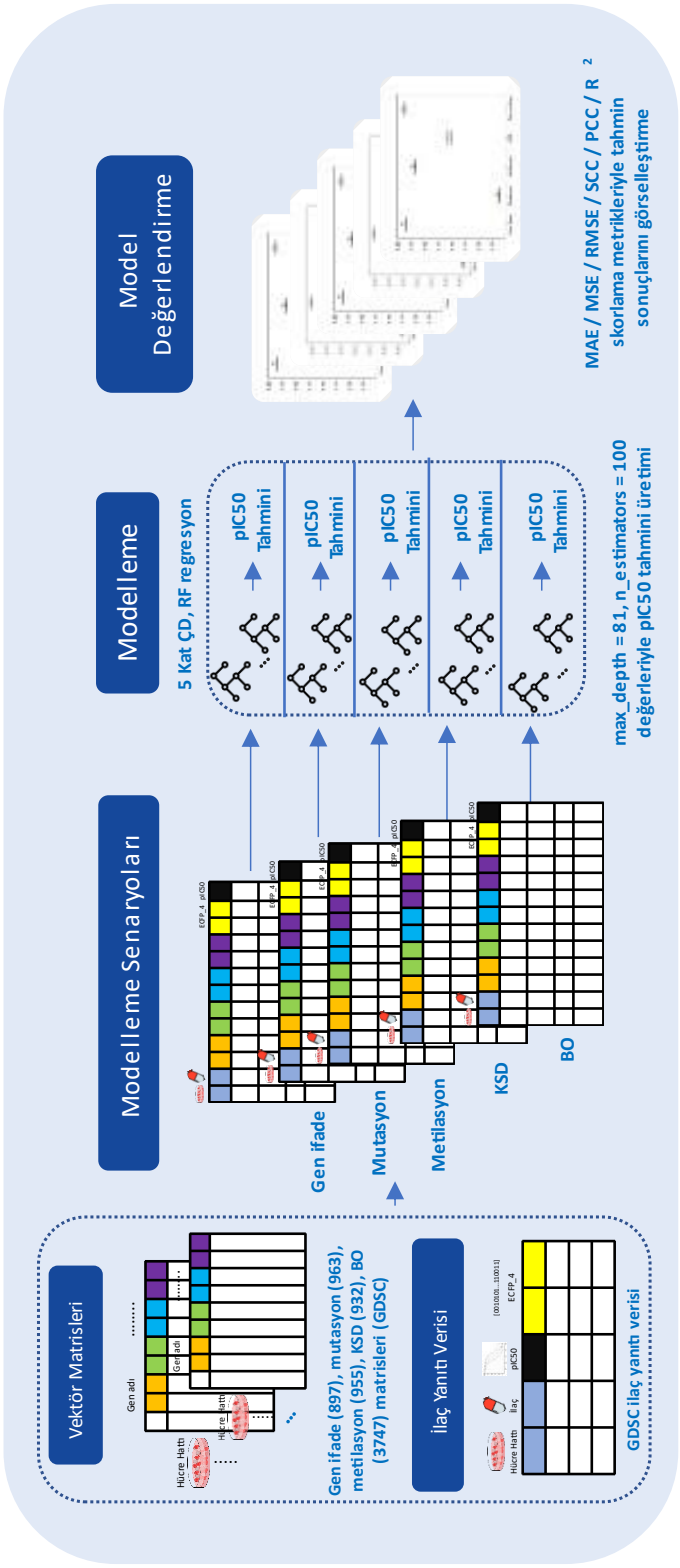
veri (ilaç yanıtı tablosu hariç) değişkenlik gösterdiğinden sadece gen ifade verisi üzerinden yapılan uygulama anlatılmıştır. Geniş ve detaylı anlatım EK-1'de yapılmıştır.

**Tablo 3.35.** Ablasyon analizi için kullanılan veri tipleri ve özellikleri.

| Veri tipi | Gen sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|
| Gen ifade | 897 | 988 | 988, 899 |
| Mutasyon | 963 | 988 | 988, 964 |
| Metilasyon | 955 | 988 | 988, 956 |
| KSD | 932 | 988 | 988, 933 |
| Birleştirilmiş omikler | 3747 | 988 | 988, 3748 |
| İlaç yanıtı | - | 988 | 337761, 6 |

Veri düzenleme aşamalarıyla modelde girdi verisi olarak kullanılacak tabloyu oluşturmak hedeflenmiştir. Bu amaçla, ilaç yanıtı verisindeki hücre hattı, ilaç ismi ve pIC50 sütunları baz alınarak düzenleme yapılmıştır. Öncelikle, 1024 bit uzunluğundaki ilaç parmak izlerinin her biri 1024 ayrı sütuna aktarılmış ve yeni bir tabloda tutulmuştur. Ardından, ilaç yanıtı verisi ve gen ifade vektör matrisi hücre hattı ismi sütunu üzerinden birleştirilerek oluşturulan yeni tabloyla dikey eksende ilaç parmak izi sütunlarına ait tablo eklenmiştir. Sonuç olarak, modelde girdi verisi olarak değerlendirilecek hücre hattı – ilaç çifti, 1903 (897 vektör uzunluğu ve 1024 parmak izi değeri toplamı) uzunluğundaki vektör ile ifade edilmiştir.

Vektör uzunluğu diğer veri tipleri değiştiği için şu sonuç vektör uzunlukları şekilde hesaplanmıştır; 1987 mutasyon vektörü (963 + 1024), 1979 metilasyon vektörü (955 + 1024), 1956 KSD vektörü (932 + 1024), 4771 birleştirilmiş veri vektörü (3747 + 1024).




**Şekil 3.7.** GDSC verisiyle yapılan ablasyon analizi için uygulanan genel işlem akışı.

RF tahmin edicisinin hazırlanmasında RandomForestRegressor metodu, max_depth (atanan 81 değeriyle) ve n_estimators (atanan 100 değeriyle) hiperparametreleriyle kullanılmıştır. 5 Kat çapraz doğrulama uygulanacağı için cross_validate metodu, RF tahmin edicisi, MAE, MSE, RMSE, SCC, PCC, $R^2$ skorlama metrikleriyle beraber kullanılmıştır. Elde edilen sonuçlar bir dosyaya yazdırılarak kaydedilmiştir. Sonuç grafikleri her metrik için ayrı şekilde çizdirilmiştir.

### 3.11.2. Veri Içi Alan Analizi

GDSC verileriyle gerçekleştirilen veri içi alan analizinde (within domain analysis, VİAA), farklı amaçları olan üç farklı değerlendirme senaryoları oluşturulmuştur. Bunlardan ilki olan hücre hattı özdeşliği temelli bölümlendirme (HHÖTB), girdi verisini hücre hattı özdeşliği üzerinden bölümleyip model eğitiminde bulunmayan hücre hatları için tahmin verme durumunun incelenmesini içerir. İkincisi, ilaç özdeşliği temelli bölümlendirme (İÖTB), ilk senaryonun ilaçlar üzerinden uygulanmasına dayanır ve eğitilen modelde tanıtılmayan ilaçların hali hazırda var olan hücre hatları üzerinde oluşturacağı ilaç yanıtının tahmini yapılır. Üçüncü senaryo olan rastgele bölümlendirmede (RB) ise, veri bütününü rastgele şekilde bölümlendirme yoluna gidilerek model eğitiminde kullanılmamış olan hücre hattı – ilaç çiftleri için ilaç yanıtı tahmini oluşturulmuştur. Şekil 3.8.'de veri içi alan analizinde uygulanan işlemlerin genelleyici görselleştirilmesi yapılmıştır.

Veri içi alan analizinde kullanılacak olan veri büyüklüğü düzenlenemeyecek boyutlarda olduğu için, verinin doku bazlı olarak düzenlendikten sonra her doku özelinde üç senaryonun değerlendirilmesi uygun görülmüştür. GDSC'deki 986 hücre hattına karşılık gelen doku isimleri kullanılarak toplamda 13 farklı doku dosyası oluşturulmuştur. Oluşturulan doku dosyaları ve genel içeriklerine dair bilgiler Tablo 3.36.'da belirtilmiştir. Gerçekleştirilen analiz aşamalarının geniş ve ayrıntılı anlatımı EK-2'de yapılmıştır.

**Tablo 3.36.** Veri içi alan analizinde kullanılmak üzere oluşturulan doku temelli dosyaların genel özellikleri.

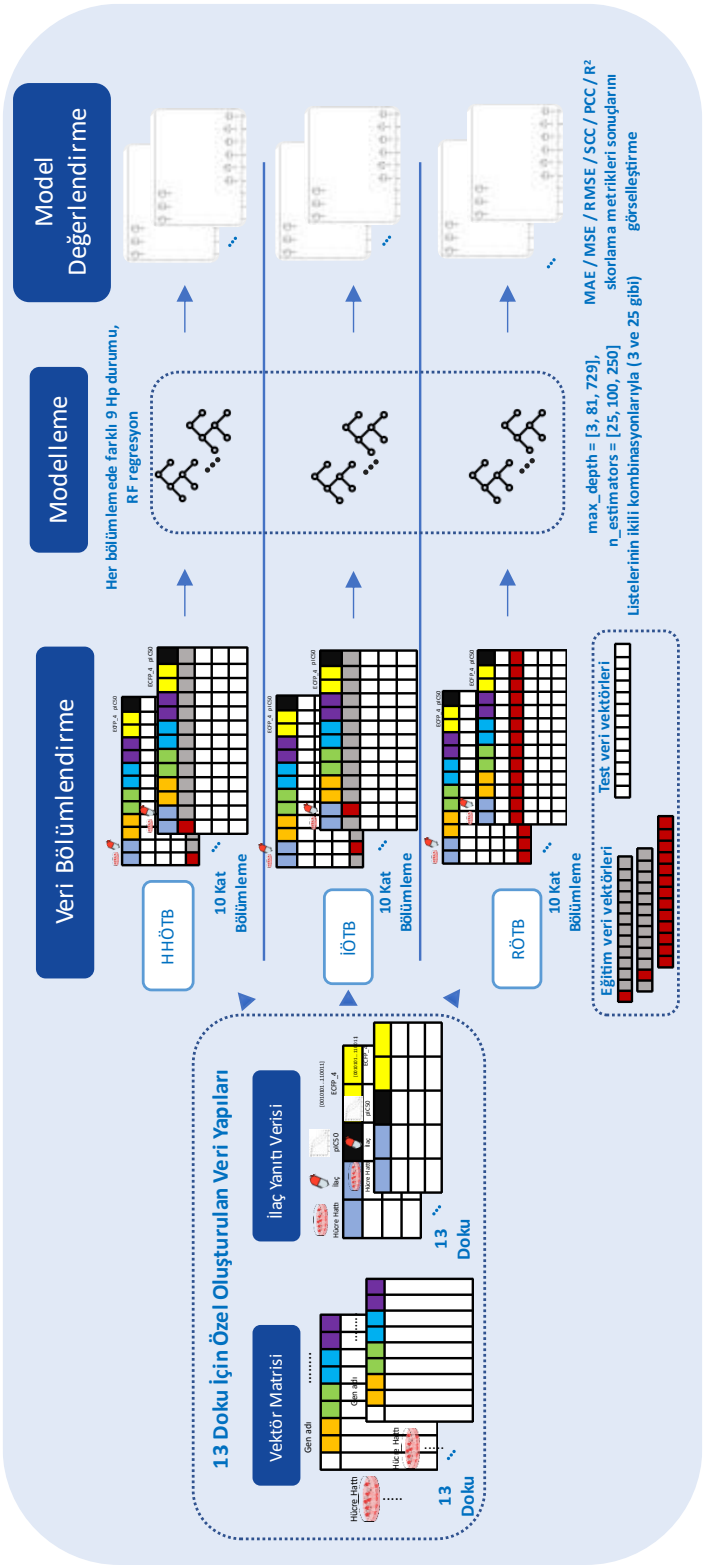
| Doku Adı | İçerdiği toplam hücre hattı sayısı | Oluşturulan tablo boyutu |
|---|---|---|
| Akciğer | 196 | 196, 1989 |
| Böbrek | 34 | 34, 1989 |
| Deri | 58 | 58, 1989 |
| Kan | 169 | 169, 1989 |
| Kemik | 39 | 39, 1989 |
| Meme | 52 | 52, 1989 |
| Pankreas | 31 | 31, 1989 |
| Sindirim sistemi | 99 | 99, 1989 |
| Sinir sistemi | 88 | 88, 1989 |
| Solunum-sindirim sistemi | 79 | 16, 1989 |
| Tiroid | 16 | 16, 1989 |
| Ürogenital sistem | 104 | 104, 1989 |
| Yumuşak doku | 21 | 21, 1989 |

Üç senaryo için ayrı ayrı ilerletilen model girdi verilerinin hazırlanma süreçlerinde benzer ön işleme adımları izlenmiştir. Girdi verisinin düzenlenmesindeki temel hedef, hücre hattı özellik vektörü ve ilaç tanımlayıcı parmak izlerinin, her ilaç yanıtı veri noktası için bir araya getirilmesidir. Bunun için, hücre hattı vektörlerini içeren doku temelli dosya ve ilaç yanıtı verisi birleştirilecektir. Kullanılan vektör dosyasındaki hücre hatları, ilaç yanıtı verisinden filtrelenir ve elde edilen 1024 bit uzunluğundaki ilaç parmak izlerindeki her sayı bir sütun hücresine yerleştirilip ilaç parmak izi tablosu oluşturulmuştur. Oluşturulan bu tabloya hücre hattı, ilaç isimleri ve pIC50 sütunları tekrar eklenmiştir. Vektör tablosu ve ilaç parmak izi tabloları hücre hattı sütunu baz alınarak birleştirilmiştir. Böylece, modelde girdi olarak kullanılacak 3013 (1989 hücre hattı özellik vektörü, 1024 ilaç parmak izi vektörü birleşimi) elemanlı vektörlerden oluşan vektör matrisi hazırlanmıştır.

RF tahmin edicisinin hazırlanmasında RandomForestRegressor metodu, max_depth (atanan 3, 81, 729 değerleriyle) ve n_estimators (atanan 25, 100, 250

değerleriyle) hiperparametreleriyle kullanılmıştır. RB senaryosu için 10 Kat çapraz doğrulama uygulanacağı için cross_validate metodu, RF tahmin edicisi, MAE, MSE, RMSE, SCC, PCC, $R^2$ skorlama metrikleriyle beraber kullanılmıştır. HHÖTB senaryosunda, analiz edilen dokuya ait hücre hattı listesi üzerinden bölümleme yapılarak, listedeki isimlerin % 90'ı model eğitimi için; kalan % 10'u ise test verisi olarak şekilde ayrılmıştır. İÖTB'de de benzer şekilde, analiz edilen doku hücre hatlarının üzerinde denenen ilaçların özgün isim listesi üzerinden % 90 / %10 bölümlendirmesi yapılmıştır. Bu bölümlendirme işlemi, özgün listelerde kayan pencere yöntemiyle 10 farklı şekilde oluşturulmuştur. RB'de ise veri 10 kat çapraz doğrulama yöntemiyle analiz edilmiştir. HHÖTB ve İÖTB'de oluşturulan 10 eğitim-test liste çifti için; RB'de uygulanan her kat çapraz doğrulama durumu için 9 hiperparametre kombinasyonu (3 max_depth ve 3 n_estimators değeri) denenmiştir. Böylece, tüm senaryolar için toplamda 90 farklı model analizi gerçekleştirilerek tahmin performansları ölçülmüştür. Ayrıca, her modelin oluşturduğu ilaç yanıtı tahminleri de kaydedilmiştir. Her skorlama metriği için sonuçların ayrı şekilde kutu grafiği çizdirilmiştir.

Üç senaryodan elde edilen sonuçlarının karşılaştırılması sonucunda, max_depth için atanan 81, n_estimators için atanan 100 değerlerinin bir arada kullanılmasının daha iyi ilaç yanıtı tahmini performansı sağladığı görülmüştür. Yapılacak olan çapraz alan ve ablasyon analizi senaryolarında da bu hiperparametreler için belirtilen değerler kullanılmıştır.




**Şekil 3.8.** GDSC verisiyle yapılan veri içi alan analizi için uygulanan genel işlem akışı.

### 3.11.3. Çapraz Alan Analizi

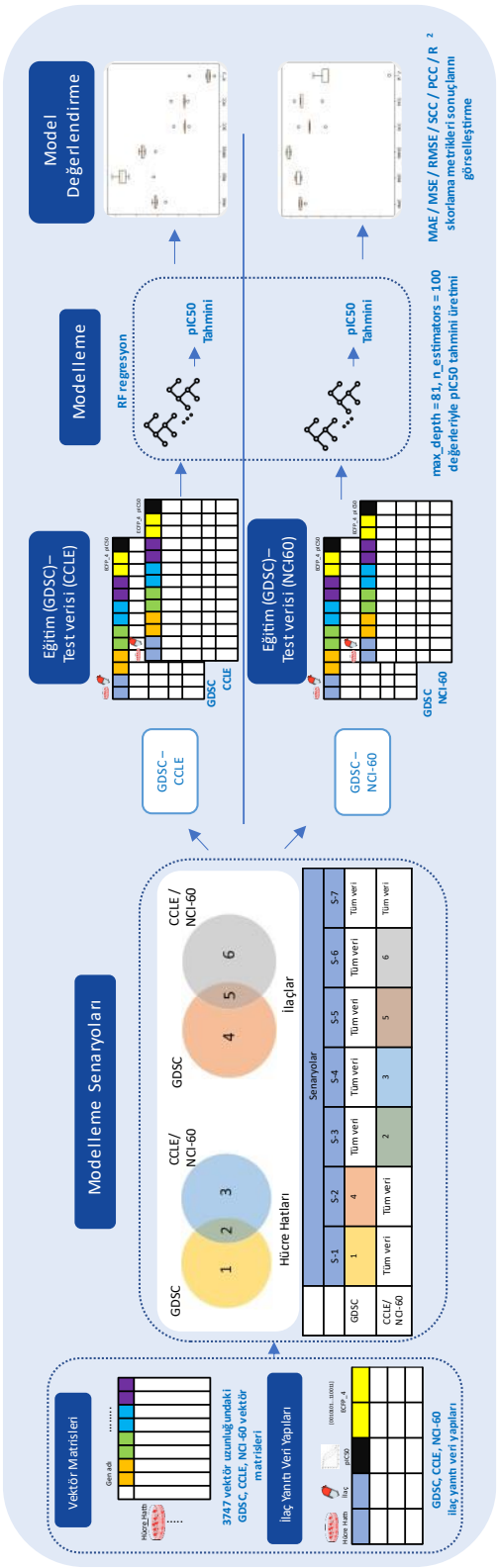
Çapraz alan analizi, belirli bir hücre hattı panel verisiyle eğitilen model üzerinde tahmin verilecek test verisi olarak başka bir panele ait veri kullanıldığı duruma karşılık gelmektedir. Bu tez çalışması kapsamında, GDSC verisi model eğitiminde kullanılmış olup test için CCLE ve NCI-60 veri setleri için ayrı şekilde ilaç yanıtı tahminleri oluşturulmuştur (Tablo 3.37.). Şekil 3.9.'da çapraz alan analizinde uygulanan işlemlerin genelleyici görselleştirilmesi yapılmıştır.

**Tablo 3.37.** Çapraz alan analizi için tasarlanan modelleme senaryoları.

| | **GDSC (Eğitim verisi)** | | **CCLE / NCI-60 (Test verisi)** | |
|---|---|---|---|---|
| **Senaryo** | **Hücre Hattı Vektör Dosyası** | **İlaç Yanıtı Vektör Dosyası** | **Hücre Hattı Dosyası** | **İlaç Yanıtı Dosyası** |
| **1** | Ortak olmayan hücre hattı | Ortak olmayan hücre hattı | Tümü | Tümü |
| **2** | Ortak olmayan ilaç | Ortak olmayan ilaç | Tümü | Tümü |
| **3** | Tümü | Tümü | Ortak hücre hattı | Ortak hücre hattı |
| **4** | Tümü | Tümü | Ortak olmayan hücre hattı | Ortak olmayan hücre hattı |
| **5** | Tümü | Tümü | Ortak ilaç | Ortak ilaç |
| **6** | Tümü | Tümü | Ortak olmayan ilaç | Ortak olmayan ilaç |
| **7** | Tümü | Tümü | Tümü | Tümü |

Bu senayolar üzerinden GDSC – CCLE arası ve GDSC – NCI-60 arası çapraz alan analizi için 7'şer model oluşturulmuştur. Tüm senaryolar için GDSC verisi modelin eğitiminde kullanılmış olup test için CCLE ve NCI-60 verileri ayrılmış aşamalarda kullanılmıştır.

Model üzerinde kullanılacak eğitim ve test verileri arasında farklı modelleme senaryoları kurgulanmıştır. İki veri arasındaki hücre hattı ve ilaç ismi ortaklıkları göz önüne alınarak oluşturulan farklı 7 durum için ilaç yanıtı tahmini yapılmıştır. Senaryolarda kullanılacak verilerin oluşturulması aşamalarında hücre hattı özellik vektörü ve ilaç yanıtı dosyaları kullanılmıştır. Bu dosyaların düzenlenme aşamaları ayrıntılı olarak EK-3'te belirtilmiştir.




**Şekil 3.9.** GDSC verisiyle yapılan çapraz alan analizi için uygulanan genel işlem akışı.

Çapraz alan analizindeki modelleme aşamalarında eğitim verisinin aynı, test verisinin farklı olduğu GDSC - CCLE arası ve GDSC - NCI-60 arası analizler iki ayrı kol olarak ilerletilmiştir ve belirlenen 7 senaryo her iki kolda da aynı şekilde uygulanmıştır. Modelleme için önceki aşamalarda uygulanan performans analizleri sonucu daha iyi tahmin performansı sunan hiperparametre değerleri kullanılmıştır (max_depth = 81 ve n_estimators = 100).

**Veri Düzenleme Aşamaları**

Yukarıda belirtilen veri seçim ve düzenleme aşamalarından sonra her bir platform için oluşturulacak olan hücre hattı özellik vektörü tablosunun toplam gen sayısı bakımından aynı hale getirilmesi için düzenlemeler yapılmıştır. Bu düzenlemeler, önceki aşamalarda oluşturulan genişletilmiş formattaki veri tiplerinden elde edilen gen listeleri için ayrı ayrı gerçekleştirilip ortak gen isimleri belirlenmiştir.

Hücre hattı ve ilaç isimlerinin tüm platformlarda aynı formatta olması sağlanmıştır. Bu format kapsamında, harf ve sayı dışındaki karakterler silinip isimler küçük harflerle değiştirilmiştir. Eğer birden çok ismin formatlı hali aynı ise asıl isimler korunup sadece küçük harflerle yazdırılmıştır.

Ekler kısmında ayrıntılı olarak belirtilen hücre hattı özellik vektörü oluşturma aşamalarında L1000 gen listesi ve ortak gen listeleri beraber kullanılmıştır. Her özellik tipi için ayrı ayrı filtreleme aşaması uygulanmıştır. İlk filtreleme aşamasında veri tablosunda gen temelli doluluk oranı hesaplanmış ve özellik tipine göre gen temelli ortalama (gen ifade, metilasyon) ve medyan değeri (KSD) hesaplanıp tablodaki ilgili tablo hücreleri doldurulmuştur. Mutasyon verisinde sadece 0 ve 1 değerleri olduğundan dolayı boşluk değerli yerler 0 değeri ile doldurulmuştur. Bir platforma özel hücre hattı vektör tablosunun üretilmesi için her özellik tipi tablosundan ilgili sütundaki değerler alınıp gen ifade, mutasyon, metilasyon, KSD sırası gözetilerek bir liste içine alınıp vektör tablosundaki hücre hattına ait satıra yatay olarak yazdırılmıştır. Her hücre hattı için

belirlenen 897 gen ifade, 963 mutasyon, 955 metilasyon, 932 KSD geni bulunduğundan, bir vektörün toplam uzunluğu 3747 olarak hesaplanmıştır.

### Çapraz Alan Analizi Için Seçilen Veri Tipleri Ve Özellikleri

Çapraz alan analizi özelinde tüm platformların hücre hattı özellik veri tipleri kendi aralarında değerlendirilip uygun olanları ile beraber özellik vektörleri oluşturulmuştur. Veri tabanları olarak hep orijinal kaynaklar kullanılmamış olup CellMiner ve DepMap sitelerinden alınan verilerden de yararlanılmıştır (Tablo 3.38.).

**Tablo 3.38.** Çapraz alan analizi için seçilen verilerin indirildiği kaynaklar.

| Panel adı | Gen ifade | Mutasyon | Metilasyon | KSD | İlaç yanıtı |
|---|---|---|---|---|---|
| GDSC | Orijinal | Orijinal | CellMiner | Orijinal (PICNIC) | Orijinal |
| CCLE | CellMiner | Orijinal | CellMiner | DepMap (ön işlem yapılmış veri) | Orijinal |
| NCI-60 | Orijinal (xai kodlu veri) | (29) | Orijinal | Orijinal (aCGH Agilent 44K) | Orijinal |

### Gen İfade Veri Tipi

Tüm platformlar için benzer aralıklarda benzer dağılıma sahip olmaları gözetilerek çeşitli kaynaklardan gen ifade verileri elde edilmiştir. Karşılaştırmalar sonucunda GDSC için orijinal; CCLE için CellMiner ve NCI-60 için orijinal veri kaynağı benzerlikleri nedeniyle tercih edilmiştir. Tablo 3.39.'da verilerin elde edilmesinde kullanılan kaynaklar ve içerikler hakkında bilgi verilmiştir.

**Tablo 3.39.** Gen ifade verisi için kaynak dosyadan üretilen veri yapılarının özellikleri.

| Panel adı | Kaynak | Gen sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|---|
| GDSC | Orijinal | 17419 | 1014 | 17419, 1015 |
| CCLE | CellMiner | 19851 | 1088 | 19851, 1089 |
| NCI-60 | Orijinal (xai kodlu veri) | 23059 | 60 | 23059, 61 |

**Mutasyon Veri Tipi**

Tüm platformlar için benzer aralıklarda benzer dağılıma sahip olmaları gözetilerek çeşitli kaynaklardan mutasyon verileri elde edilmiştir. Karşılaştırmalar sonucunda GDSC için orijinal; CCLE için orijinal ve NCI-60 için Chen ve Zhang'in (29) yaptığı araştırmanın Ek kısmında belirtilen dosya, veri kaynağı benzerlikleri nedeniyle tercih edilmiştir. Tablo 3.40.'ta verilerin elde edilmesinde kullanılan kaynaklar ve içerikler hakkında bilgi verilmiştir.

**Tablo 3.40.** Mutasyon verisi için kaynak dosyadan üretilen veri yapılarının özellikleri.

| Panel adı | Kaynak | Gen sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|---|
| GDSC | Orijinal | 21972 | 1032 | 21972, 1033 |
| CCLE | Orijinal | 19286 | 1570 | 19286, 1571 |
| NCI-60 | (29) | 443 | 59 | 443, 60 |

**Metilasyon Veri Tipi**

Tüm platformlar için benzer aralıklarda benzer dağılıma sahip olmaları gözetilerek çeşitli kaynaklardan mutasyon verileri elde edilmiştir. Karşılaştırmalar sonucunda GDSC için orijinal; CCLE için orijinal ve NCI-60 için veri kaynağı benzerlikleri nedeniyle tercih edilmiştir. Tablo 3.41.'de verilerin elde edilmesinde kullanılan kaynaklar ve içerikler hakkında bilgi verilmiştir.

**Tablo 3.41.** Metilasyon verisi için kaynak dosyadan üretilen veri yapılarının özellikleri.

| Panel Adı | Kaynak | Gen sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|---|
| GDSC | CellMiner | 1080 | 19864 | 19864, 1081 |
| CCLE | CellMiner | 19880 | 1089 | 19880, 1090 |
| NCI-60 | Orijinal | 17552 | 60 | 17552, 61 |

**KSD Veri Tipi**

Tüm platformlar için benzer aralıklarda benzer dağılıma sahip olmaları gözetilerek çeşitli kaynaklardan mutasyon verileri elde edilmiştir. Karşılaştırmalar sonucunda GDSC

için orijinal; CCLE için DepMap ve NCI-60 orijinal veri kaynağı benzerlikleri nedeniyle tercih edilmiştir. Tablo 3.42.'de verilerin elde edilmesinde kullanılan kaynaklar ve içerikler hakkında bilgi verilmiştir.

**Tablo 3.42.** KSD verisi için kaynak dosyadan üretilen veri yapılarının özellikleri.

| Panel Adı | Kaynak | Gen sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|---|
| GDSC | Orijinal (PICNIC) | 2450 | 986 | 24502, 987 |
| CCLE | DepMap (ön işlem yapılmış veri) | 25368 | 1754 | 25368, 1755 |
| NCI-60 | Orijinal (aCGH Agilent 44K) | 19847 | 60 | 19905, 61 |

**İlaç Yanıtı Veri Tipi**

GDSC, CCLE, NCI-60 için orijinal ilaç yanıtı verileri kullanılmıştır. Tüm platformlar için benzer aralıklarda benzer dağılıma sahip olmaları gözetilerek çeşitli kaynaklardan mutasyon verileri elde edilmiştir. Karşılaştırmalar sonucunda GDSC için orijinal; CCLE için orijinal ve NCI-60 için orijinal veri kaynağı benzerlikleri nedeniyle tercih edilmiştir. Tablo 3.43'te verilerin elde edilmesinde kullanılan kaynaklar ve içerikler hakkında bilgi verilmiştir.

**Tablo 3.43.** İlaç yanıtı verisi için kaynak dosyadan üretilen tabloların özellikleri.

| Panel Adı | Kaynak | İlaç sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|---|
| GDSC | Orijinal | 404 | 988 | 337761, 6 |
| CCLE | Orijinal | 24 | 504 | 11670, 6 |
| NCI-60 | Orijinal | 1054 | 60 | 60678, 7 |

**Çapraz Alan Analizi İçin Kullanılan Veri Yapıları ve Uygulanan Modelleme Senaryoları**

Tablo 3.44.'te, modelleme senaryolarında kullanılacak olan omik veri tiplerindeki genler için ortaklıklar bulunmadan önceki durum sunulmuştur. Tablo 3.45'te ise platformlar arası omik veri tipi temelli olarak ortak genler bulunduktan sonra kalan gen sayıları gösterilmektedir.

**Tablo 3.44.** Tüm platformlarda hücre hatlarına ait vektörlerin oluşturulmasında kullanılacak olan ortak genlerin her bir hücre hattı özellik tipi bazında hesaplanması.

| Veri tipi | GDSC Gen Sayısı | CCLE Gen Sayısı | NCI-60 Gen Sayısı | Ortak Gen Sayısı |
|---|---|---|---|---|
| Gen ifade | 923 | 960 | 951 | 897 |
| Mutasyon | 978 | 970 | 963 | 963 |
| Metilasyon | 973 | 964 | 963 | 955 |
| KSD | 945 | 934 | 963 | 932 |

**Tablo 3.45.** Platformlara ait hücre hattı özellik vektörü verilerinin düzenlenmesinde kullanılan gen temelli sayılar ve ilgili vektör tablosunda bulunan hücre hattı sayısı.

| Panel Adı | Hücre hattı sayısı | Gen ifade gen sayısı | Mutasyon gen sayısı | Metilasyon gen sayısı | KSD gen sayısı |
|---|---|---|---|---|---|
| GDSC | 988 | 897 | 963 | 955 | 932 |
| CCLE | 503 | 897 | 963 | 955 | 932 |
| NCI-60 | 60 | 897 | 963 | 955 | 932 |

Oluşturulan modeller joblib paketi yardımıyla kaydedilmiştir. MAE, MSE, RMSE, SCC, PCC, $R^2$ metrikleri ile beraber modelin performansı ölçülerek kaydedilmiştir. Gerçek değerler ve tahmin değerleri ilgili ilaç – hücre hattı çifti isimleri ile beraber kaydedilmiştir. Analizlerin ayrıntılı aşamaları EK-3'te belirtilmiştir.

### 3.12. DeepResponse-RF ile *In Vitro*'da Doğrulaması Yapılacak İlaç Yanıtı Tahminlerinin Üretilmesi

Yöntemimizin GDSC verisi üzerinden üreteceği ilaç yanıtı tahminleri için *in vitro* doğrulamaları, dahil olduğumuz TÜSEB projesinde ortağımız olan Orta Doğu Teknik Üniversitesi (ODTÜ) bünyesindeki KanSiL Lab tarafından yapılmıştır. Bunun için öncelikle, KanSiL Lab envanterinde ve GDSC verisinde ortak olan hücre hatları seçilmiştir. İlaç molekülleri için ise böyle bir ayrım yapılmayarak tüm ilaçlara ait parmak izleri kullanılmıştır. DeepResponse-RF ile 12 ortak hücre hattı ve 35 ilaç molekülünden oluşan toplamda 420 hücre hattı – ilaç çiftine yönelik ilaç yanıtı tahminleri oluşturulmuştur.

Ortak hücre hatlarının çoğunun karaciğer organına ait olması nedeniyle modelde eğitim veri seti olarak sindirim sistemi dokusuna ait veri seti kullanılmıştır. Modelde kullanılacak GDSC'ye ait veri seti düzenlenmesinde dikey olarak birleştirilmiş özellik vektörü, parmak izleri sütunları ve pIC50 değeri sütunu ve 34557 satır (hücre hattı-ilaç çifti) bulunmaktadır. Hücre hattına ait özellik vektörü, 35542 tane sütun, ilaç moleküllerine ait parmak izi için 1024 tane sütun ve pIC50 değerleri için 1 adet sütun olmak üzere toplamda 36566 sütundan oluşmaktadır. Benzer düzenle oluşturulan ve üzerinde tahmin verilen KanSiL Lab verisinde ise sadece pIC50 sütunu bulunmamaktadır.

DeepResponse-RF modeli için iki tane hiperparametre belirlenmiştir. Oluşturulan karar ağaçlarının maksimum derinliğini belirtmekte *max_depth* hiperparametresi için değer olarak "81" değeri atanmıştır. Karar ağaçlarının oluşturduğu orman içindeki maksimum ağaç sayısını belirleyen *n_estimators* hiperparametresi için ise değer olarak "100" değeri atanmıştır. Atanan değerler performans ölçümlerinde kullanılan metriklerin gösterdiği hata oranlarındaki değişimler dikkate alınarak verilmiştir.

Model yardımıyla tahmin oluşturulan KanSiL Lab'a ait 420 hücre hattı-ilaç çiftinden 100 tanesinin GDSC'de pIC50 ilaç yanıtı karşılığı olduğu görülmüştür. Geriye kalan 320 çift ise GDSC'de eşleşmesi olmayanlar olarak kaydedilmiştir. 420 çiftin tümüne ait sonuçlar ise EK-4'te (Tablo 8.23.) sunulmuştur.

Yukarıda bahsedilen 420 çift içinden in vitro doğrulama deneyleri yapılacak olan çiftlerin belirlenmesi için bir prosedür izlenmiştir. KanSiL Lab karaciğer kanseri üzerinde çalıştığı için, in vitro deneylerde kullanılmak üzere seçilen hücre hatları da hepatosellüler karsinom hücre hatları olmuştur. Prosedürdeki ilk aşamada, hem GDSC üzerinde hem literatürde deneysel verisi olmayan çiftler filtrelenmiştir. Ardından, kalan hücre hattı ve ilaçların ayrı ayrı şekilde deneysel veri noktaları incelenip, bunların verilen tahminlerle arasındaki farkın az olmasına dikkat edilerek filtreleme yapılmıştır. Yapılan bu ikinci filtreleme, o hücre hattı veya ilaç için güvenilir sonuçlar üretildiğinin desteklenmesi amacıyla yapılmıştır.

### 3.13. DrugBank Kaynaklı İlaçlar İçin GDSC Verisi Üzerinden İlaç Yanıtı Tahminleri Üretilmesi

DrugBank veri tabanından indirilen kaynak dosyada SMILES notasyonu bulunan 10739 ilaç için parmak izleri RDkit paketi kullanılarak oluşturulmuştur. GDSC sindirim sistemi verisinde bulunan 99 hücre hattı ile bu ilaçların tüm kombinasyonları için 1063161 tahmin verilecek hücre hattı – ilaç çifti belirlenmiştir.

Belirlenen çiftler için veri, önceki yapılara benzer şekildedir. Bir çift için önce GDSC hücre hattına ait uzunluğu 35542 olan özellik vektörü; sonra ise DrugBank ilacına ait uzunluğu 1024 olan parmak izi vektörü sütunlara ayrılmış olarak belirtilmiştir (bir çift için toplamda 36566 ayrı sütun değeri). Oluşturulan veri, hücre hattı temelli olarak ayrıma tutulup önceden oluşturulmuş DeepResponse-RF modeline verilerek tahminler elde edilmiştir. Son adımda, tüm çiftlere ait tahminler birleştirilmiştir. Sonuç tablosunun en yüksek pIC50 değerli 100 çifti EK-4'te (Tablo 8.25.) sunulmuştur.

# 4. BULGULAR

## 4.1. Veri Araştırması

Gerçekleştirilen analizlerde kullanılan veri yapılarının genel özellikleri bu başlık altında incelenmiştir.

### 4.1.1. Ablasyon Analizi

Ablasyon analizi kapsamında değerlendirilen veri yapılarının içeriklerine dair genel özellikler Tablo 4.1.'de verilmiştir.

**Tablo 4.1.** GDSC ablasyon analizinde kullanılan veri tipleri için oluşturulan dosyaların genel özellikleri.

| Veri tipi | Gen sayısı | Hücre hattı sayısı | İlaç sayısı | Tablo boyutu |
|---|---|---|---|---|
| Gen ifade | 897 | 988 | - | 988, 899 |
| Mutasyon | 963 | 988 | - | 988, 964 |
| Metilasyon | 955 | 988 | - | 988, 956 |
| KSD | 932 | 988 | - | 988, 933 |
| Birleştirilmiş omikler | 3747 | 988 | - | 988, 3748 |
| İlaç yanıtı | - | 988 | 404 | 337761, 6 |

### 4.1.2. Veri İçi Alan Analizi

GDSC verisi üzerinden hazırlanmış olan omik veri tiplerinin her biri için değerlendirilen veri yapılarının genel özellikleri aşağıda belirtilmiştir (Tablo 4.2. – 4.5.).

**Tablo 4.2.** GDSC Gen ifade veri tipi (VİAA için) için oluşturulan dosyaların genel özellikleri.

| Veri tipi | Gen sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|
| GDSC gen ifade kaynak veri | 17737 | 1018 | 17737, 1020 |
| GDSC gen ifade ön işlemli veri | 17419 | 1014 | 17419, 1015 |
| GDSC gen ifade genişletilmiş veri | 27794 | 1126 | 27794, 1127 |

**Tablo 4.3.** GDSC Mutasyon veri tipi (VİAA) için oluşturulan dosyaların genel özellikleri.

| Veri tipi | Gen sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|
| GDSC mutasyon kaynak veri | 21972 | 1032 | 1796526, 8 |
| GDSC mutasyon ön işlemli veri | 21972 | 1032 | 21972, 1033 |
| GDSC mutasyon genişletilmiş veri | 27794 | 1126 | 27794, 1127 |

**Tablo 4.4.** GDSC Metilasyon veri tipi (VİAA) için oluşturulan dosyaların genel özellikleri.

| Veri tipi | Gen sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|
| GDSC metilasyon kaynak veri | 378 | 790 | 378, 791 |
| GDSC metilasyon ön işlemli veri | 427 | 790 | 427, 791 |
| GDSC metilasyon genişletilmiş veri | 27794 | 1126 | 27794, 1127 |

**Tablo 4.5.** GDSC KSD veri tipi (VİAA) için oluşturulan dosyaların genel özellikleri.

| Veri tipi | Gen sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|
| GDSC KSD kaynak veri | 24502 | 986 | 24502, 987 |
| GDSC KSD ön işlemli veri | 24502 | 986 | 24502, 987 |
| GDSC KSD genişletilmiş veri | 27794 | 1126 | 27794, 1127 |

### Hücre Hattı Özellik Vektörlerinde Kullanılacak Gen Sayısının Azaltılması

Hücre hattı özellik vektörlerinde kullanılacak olan omik veri tiplerinin her biri üzerinde uygulanan eşik değerleriyle değerlendirmeye alınan gen sayısı azaltılmıştır. İki farklı boyutlu vektör tipini (35542 ve 1910) oluşturmada kullanılan eşik değerleri ve tablolardaki genel içerikler Tablo 4.6. - 4.7.'de verilmiştir. Bahsedilen iki vektör tipi ve farklı hiperparametre seçimleri yapılarak eğitilen DeepResponse-RF modellerinin karşılaştırılması Tablo 4.8.'de yapılmıştır. Karşılaştırma tablosunda 1910 uzunluklu vektör kullanılarak elde edilen sonuçların 35542 uzunluklu vektörle elde edilenlere göre hem

MAE metriğinde yüksek performanslı hem zaman açısından avantajlı olduğu görülmüştür.

**Tablo 4.6.** GDSC hücre hattı özellik vektör matrisi (VİAA – Sindirim sistemi verisi üzerinde ve L1000 filtrelemesi olmadan) oluşturmak için kullanılan veri tiplerinde boyut küçültme amacıyla uygulanan eşik değerleri.

| **Ön işleme aşaması** | **Uygulanan Eşik değeri (%)** | **Gen sayısı** | **Hücre hattı sayısı** | **Tablo boyutu** |
|---|---|---|---|---|
| Gen ifade | 90 | 16468 | 1126 | 1126, 35543 |
| Mutasyon | 5 | 2458 | 1126 | 1126, 35543 |
| Metilasyon | 1 | 148 | 1126 | 1126, 35543 |
| KSD | 97 | 16468 | 1126 | 1126, 35543 |
| **Vektör Matrisi** | **-** | **35542** | **1126** | **1126, 35543** |

**Tablo 4.7.** GDSC hücre hattı özellik vektör matrisi (VİAA – Sindirim sistemi verisi üzerinde ve L1000 filtrelemesiyle) oluşturmak için kullanılan veri tiplerinde boyut küçültme amacıyla uygulanan eşik değerleri.

| **Ön işleme aşaması** | **Uygulanan Eşik değeri (%)** | **Gen sayısı** | **Hücre hattı sayısı** | **Tablo boyutu** |
|---|---|---|---|---|
| Gen ifade | 90 | 896 | 1126 | 1126, 1911 |
| Mutasyon | 5 | 107 | 1126 | 1126, 1911 |
| Metilasyon | 1 | 11 | 1126 | 1126, 1911 |
| KSD | 97 | 896 | 1126 | 1126, 1911 |
| **Vektör Matrisi** | **-** | **1910** | **1126** | **1126, 1911** |

Yukarıda belirtilen karşılaştırmalardan sonra, veri içi alan analizinde kullanılmak üzere üretilen vektörlerde bulunacak genlerin hem eşik değerlere uygun hem L1000 listesinde olanları tercih edilmiştir. Buna bağlı olarak kullanılan omik veri tiplerinin her biri için uygulanan eşik değerleri ve ilgili tablo özellikleri Tablo 4.9.'da belirtilmiştir. GDSC verisinin doku temelli bölümlendirilmesiyle oluşan 13 dokuya ait veri yapılarının özellikleri Tablo 4.10.'da verilmiştir. Bu analiz tipinde kullanılacak ilaç yanıtı veri tipi için oluşturulan yapıların içerikleri Tablo 4.11.'de özetlenmiştir.

**Tablo 4.8.** GDSC sindirim sistemi verisinin farklı vektör uzunluklarına sahip matrisleriyle farklı hiperparametre seçimleri yapılarak eğitilen DeepResponse-RF modellerinin karşılaştırılması.

| Kullanılan veri | Hiperparametre Seçimleri | Analiz süresi (saat) | MAE skoru |
|---|---|---|---|
| Tüm sindirim sistemi vektörü (35542 uzunluklu) | "bootstrap" : [True], "max_depth" : [5], "min_samples_leaf" : [1], "min_samples_split" : [2], "n_estimators" : [100], "max_features" : ["auto"] | 30 | 0.745 |
| Tüm sindirim sistemi vektörü (35542 uzunluklu) | "bootstrap" : [True], "max_depth" : [5], "min_samples_leaf" : [1], "min_samples_split" : [2], "n_estimators" : [25], "max_features" : ["auto"], "oob_score" : [True] | 27 | 0.746 |
| Tüm sindirim sistemi vektörü (35542 uzunluklu) | "bootstrap" : [True], "max_depth" : [5], "min_samples_leaf" : [1], "min_samples_split" : [2], "n_estimators" : [100], "max_features" : ["auto"], "oob_score" : [True] | 48 | 0.746 |
| L1000 ile filtrelenen sindirim sistemi vektörü (1910 uzunluklu) | "bootstrap" : [True], "max_depth" : [5], "min_samples_leaf" : [1], "min_samples_split" : [2], "n_estimators" : [100], "max_features" : ["auto"], "oob_score" : [True] | 2 | 0.74 |
| L1000 ile filtrelenen sindirim sistemi vektörü (1910 uzunluklu) | "bootstrap" : [True], "max_depth" : [5], "min_samples_leaf" : [1], "min_samples_split" : [2], "n_estimators" : [25], "max_features" : ["auto"], "oob_score" : [True] | 0.5 | 0.74 |

**Tablo 4.9.** GDSC hücre hattı özellik vektör matrisi (VİAA – tüm dokularda ve L1000 filtrelemeyle) oluşturmak için kullanılan veri tiplerinde boyut küçültme amacıyla uygulanan eşik değerleri.

| Ön işleme aşaması | Uygulanan Eşik değeri (%) | Gen sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|---|
| Gen ifade | 90 | 923 | 986 | 923, 987 |
| Mutasyon | 5 | 110 | 986 | 110, 987 |
| Metilasyon | 1 | 12 | 986 | 12, 987 |
| KSD | 87 | 944 | 986 | 944, 987 |
| **Vektör Matrisi** | **-** | **1989** | **986** | **1126, 987** |

**Tablo 4.10.** GDSC doku temelli oluşturulan vektör matrisi dosyalarının genel özellikleri.

| Doku adı | Gen sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|
| Akciğer | 1989 | 196 | 196, 1989 |
| Böbrek | 1989 | 34 | 34, 1989 |
| Deri | 1989 | 58 | 58, 1989 |
| Meme | 1989 | 169 | 169, 1989 |
| Kan | 1989 | 39 | 39, 1989 |
| Kemik | 1989 | 52 | 52, 1989 |
| Pankreas | 1989 | 31 | 31, 1989 |
| Sindirim sistemi | 1989 | 99 | 99, 1989 |
| Sinir sistemi | 1989 | 88 | 88, 1989 |
| Solunum-sindirim sistemi | 1989 | 79 | 16, 1989 |
| Tiroid | 1989 | 16 | 16, 1989 |
| Ürogenital sistem | 1989 | 104 | 104, 1989 |
| Yumuşak doku | 1989 | 21 | 21, 1989 |

**Tablo 4.11.** GDSC ilaç yanıtı veri tipi (VİAA) için oluşturulan dosyaların genel özellikleri.

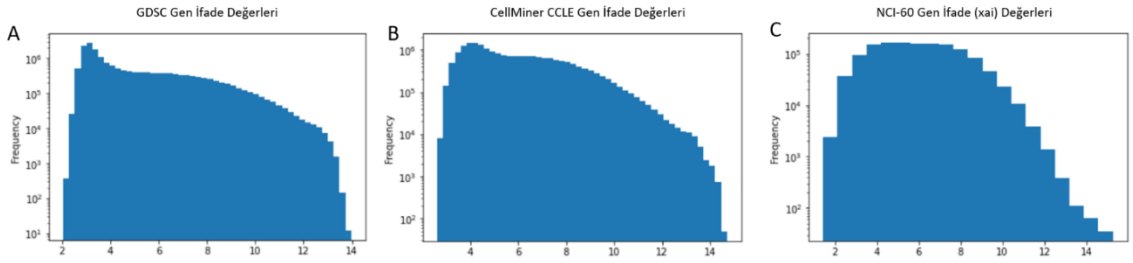
| Veri tipi | İlaç sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|
| İlaç yanıtı verisi (ilaç ismi ile) | 406 | 986 | 338950, 3 |
| İlaç yanıtı verisi (parmak izi ile) | 406 | 986 | 338950, 3 |

### 4.1.3 Çapraz Alan Analizi

Çapraz alan analizinde tüm panellerdeki omik ve ilaç yanıtı veri tiplerine ait tabloların içerikleri, kendilerine özel başlıklar altında beraberce özetlenmiştir.

### Gen İfade Veri Tipi

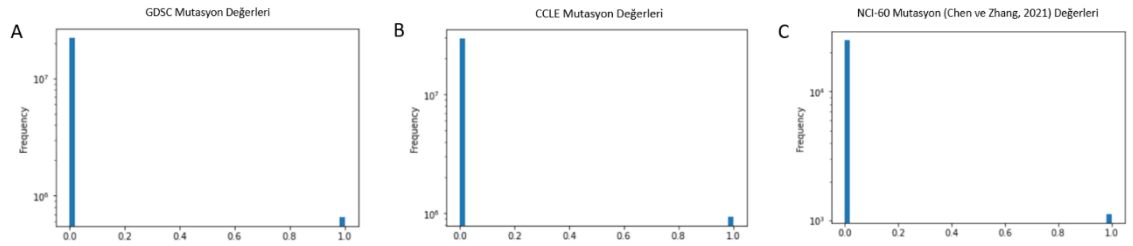
Gen ifade veri tipi için oluşturulan ön işlemli veri yapılarına ait özellikler önceki bölümde sırasıyla özetlenmiştir (Bkz. Tablo 3.39.). Şekil 4.1.'de de ön işlemli gen ifade verilerine ait histogram grafikleri verilmiştir.




**Şekil 4.1.** Tüm platformlardaki gen ifade verilerine ait histogram grafikleri. (A), GDSC; (B), CCLE; (C), NCI-60.

### Mutasyon Veri Tipi

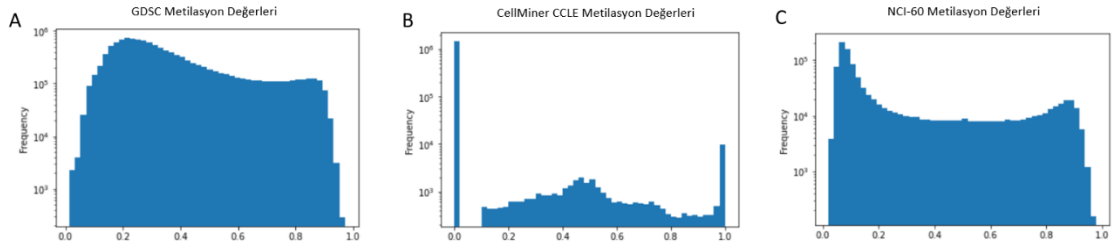
Mutasyon veri tipi için oluşturulan ön işlemli veri yapılarına ait özellikler önceki bölümde sırasıyla özetlenmiştir (Bkz. Tablo 3.40.). Şekil 4.2.'de ise ön işlemli mutasyon verilerine ait histogram grafikleri verilmiştir.




**Şekil 4.2.** Tüm platformlardaki mutasyon verilerine ait histogram grafikleri. (A), GDSC; (B), CCLE; (C), NCI-60.

### Metilasyon Veri Tipi

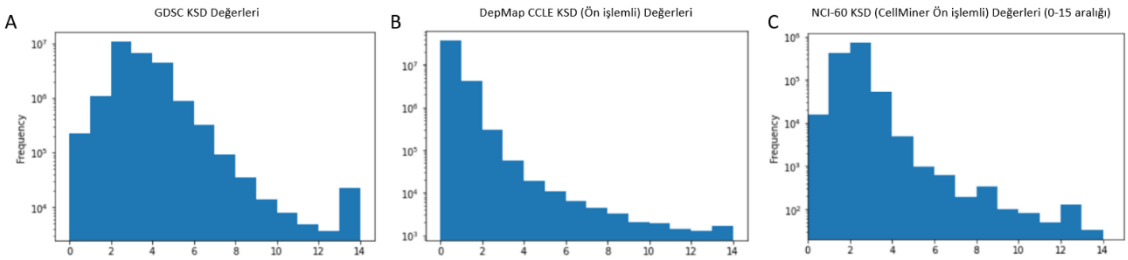
Metilasyon veri tipi için oluşturulan ön işlemli veri yapılarına ait özellikler önceki bölümde sırasıyla özetlenmiştir (Bkz. Tablo 3.41.). Ön işlemli metilasyon verilerine ait histogram grafikleri Şekil 4.3.'te verilmiştir.




**Şekil 4.3.** Tüm platformlardaki metilasyon verilerine ait histogram grafikleri. (A), GDSC; (B), CCLE; (C), NCI-60.

**KSD Veri Tipi**

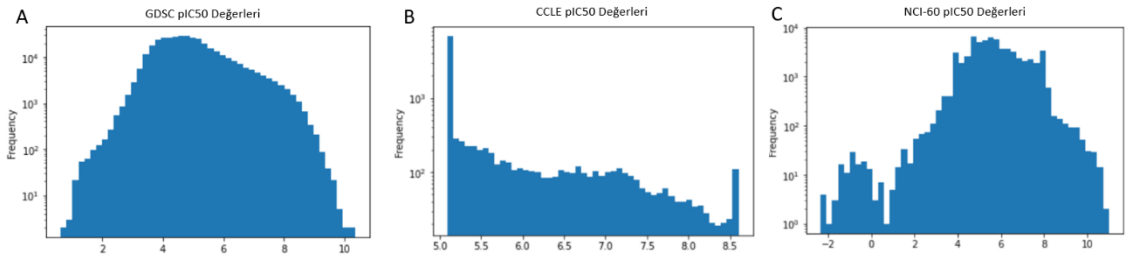
KSD veri tipi için oluşturulan ön işlemli veri yapılarına ait özellikler önceki bölümde sırasıyla özetlenmiştir (Bkz. Tablo 3.42.). Ön işlemli KSD verilerine ait histogram grafikleri ise Şekil 4.4.'te verilmiştir.




**Şekil 4.4.** Tüm platformlardaki KSD verilerine ait histogram grafikleri. (A), GDSC; (B), CCLE; (C), NCI-60.

**İlaç Yanıtı Veri Tipi**

İlaç Yanıtı veri tipi için oluşturulan ön işlemli veri yapılarına ait özellikler önceki bölümde sırasıyla özetlenmiştir (Bkz. Tablo 3.43.). Ön işlemli ilaç yanıtı veri tiplerine ait histogram grafikleri ise Şekil 4.5.'te verilmiştir.




**Şekil 4.5.** Tüm platformlardaki ilaç yanıtı verilerine ait histogram grafikleri. (A), GDSC; (B), CCLE; (C), NCI-60.

Aşağıda ayrı başlıklar altında, çapraz alan analizi için oluşturulan senaryolarda kullanılan veri yapılarının içerikleri özetlenmiştir. GDSC veri yapıları, Tablo 4.12. – 4.13.'te; CCLE'ye ait olanlar Tablo 4.14. – 4.15'te; NCI-60 için oluşturulanlar ise Tablo 4.16 - 4.17.'de belirtilmiştir.

**GDSC Verisi**

**Tablo 4.12.** ÇAA senaryolarında kullanılan GDSC hücre hattı özellik vektör matrisleri için oluşturulan dosyaların genel özellikleri.

| Senaryo | Toplam gen sayısı (vektör boyutu) | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|
| 1 (GDSC-CCLE) | 3747 | 612 | 612, 3748 |
| 1 (GDSC - NCI-60) | 3747 | 935 | 935, 3748 |
| 2 (GDSC-CCLE) | 3747 | 988 | 988, 3748 |
| 2 (GDSC - NCI-60) | 3747 | 988 | 988, 3748 |
| 3 (GDSC-CCLE) | 3747 | 988 | 988, 3748 |
| 3 (GDSC - NCI-60) | 3747 | 988 | 988, 3748 |
| 4 (GDSC-CCLE) | 3747 | 988 | 988, 3748 |
| 4 (GDSC - NCI-60) | 3747 | 988 | 988, 3748 |
| 5 (GDSC-CCLE) | 3747 | 988 | 988, 3748 |
| 5 (GDSC - NCI-60) | 3747 | 988 | 988, 3748 |
| 6 (GDSC-CCLE) | 3747 | 988 | 988, 3748 |
| 6 (GDSC - NCI-60) | 3747 | 988 | 988, 3748 |
| 7 (GDSC-CCLE) | 3747 | 988 | 988, 3748 |
| 7 (GDSC - NCI-60) | 3747 | 988 | 988, 3748 |

**Tablo 4.13.** ÇAA senaryolarında kullanılan GDSC ilaç yanıtı veri tipi dosyalarının genel özellikleri.

| Senaryo | İlaç sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|
| 1 (GDSC-CCLE) | 404 | 612 | 203934, 6 |
| 1 (GDSC-NCI-60) | 404 | 935 | 318661, 6 |
| 2 (GDSC-CCLE) | 391 | 988 | 327231, 6 |
| 2 (GDSC-NCI-60) | 307 | 988 | 255811, 6 |
| 3 (GDSC-CCLE) | 404 | 988 | 337761, 6 |
| 3 (GDSC-NCI-60) | 404 | 988 | 337761, 6 |
| 4 (GDSC-CCLE) | 404 | 988 | 337761, 6 |
| 4 (GDSC-NCI-60) | 404 | 988 | 337761, 6 |
| 5 (GDSC-CCLE) | 404 | 988 | 337761, 6 |
| 5 (GDSC-NCI-60) | 404 | 988 | 337761, 6 |
| 6 (GDSC-CCLE) | 404 | 988 | 337761, 6 |
| 6 (GDSC-NCI-60) | 404 | 988 | 337761, 6 |
| 7 (GDSC-CCLE) | 404 | 988 | 337761, 6 |
| 7 (GDSC-NCI-60) | 404 | 988 | 337761, 6 |

**CCLE Verisi**

**Tablo 4.14.** ÇAA senaryolarında kullanılan CCLE hücre hattı özellik vektör matrisleri için oluşturulan dosyaların genel özellikleri.

| Senaryo | Toplam gen sayısı (vektör boyutu) | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|
| 1 | 3747 | 503 | 503, 3748 |
| 2 | 3747 | 503 | 503, 3748 |
| 3 | 3747 | 376 | 376, 3748 |
| 4 | 3747 | 127 | 127, 3748 |
| 5 | 3747 | 503 | 503, 3748 |
| 6 | 3747 | 503 | 503, 3748 |
| 7 | 3747 | 503 | 503, 3748 |

**Tablo 4.15.** ÇAA senaryolarında kullanılan CCLE ilaç yanıtı veri tipi dosyalarının genel özellikleri.

| Senaryo | İlaç sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|
| 1 | 24 | 504 | 11670, 6 |
| 2 | 24 | 504 | 11670, 6 |
| 3 | 24 | 376 | 8666, 6 |
| 4 | 24 | 128 | 3004, 6 |
| 5 | 13 | 504 | 6265, 6 |
| 6 | 11 | 504 | 5405, 6 |
| 7 | 24 | 504 | 11670, 6 |

**NCI-60 Verisi**

**Tablo 4.16.** ÇAA senaryolarında kullanılan NCI-60 hücre hattı özellik vektör matrisleri için oluşturulan dosyaların genel özellikleri.

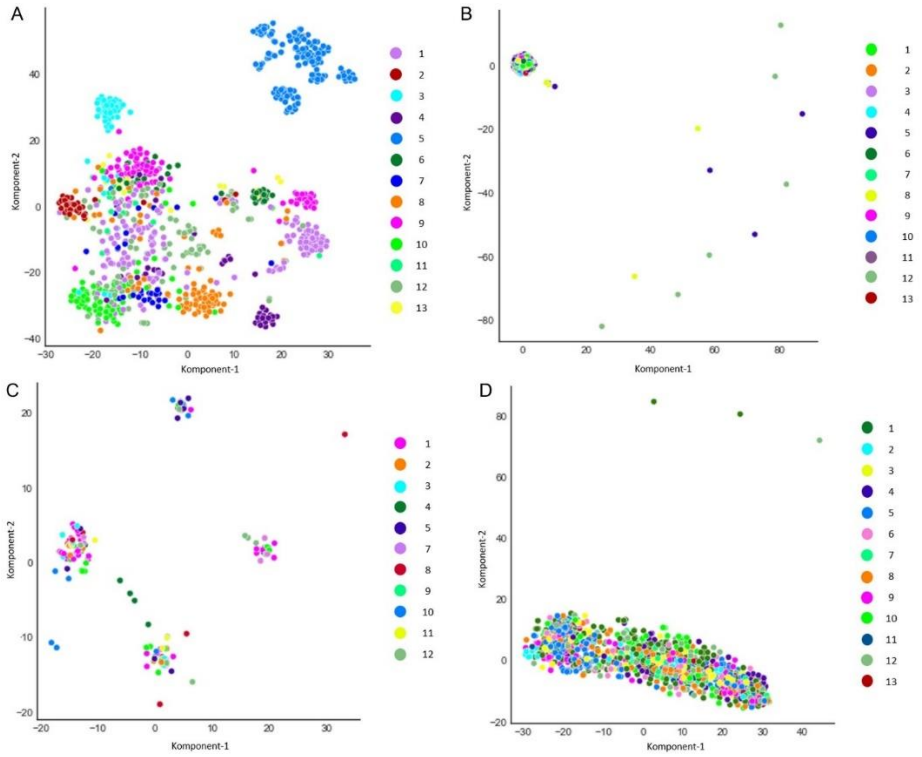
| Senaryo | Toplam gen sayısı (vektör boyutu) | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|
| 1 | 3747 | 60 | 60, 3748 |
| 2 | 3747 | 60 | 60, 3748 |
| 3 | 3747 | 53 | 53, 3748 |
| 4 | 3747 | 7 | 7, 3748 |
| 5 | 3747 | 60 | 60, 3748 |
| 6 | 3747 | 60 | 60, 3748 |
| 7 | 3747 | 60 | 60, 3748 |

**Tablo 4.17.** ÇAA senaryolarında kullanılan NCI-60 ilaç yanıtı veri tipi dosyalarının genel özellikleri.

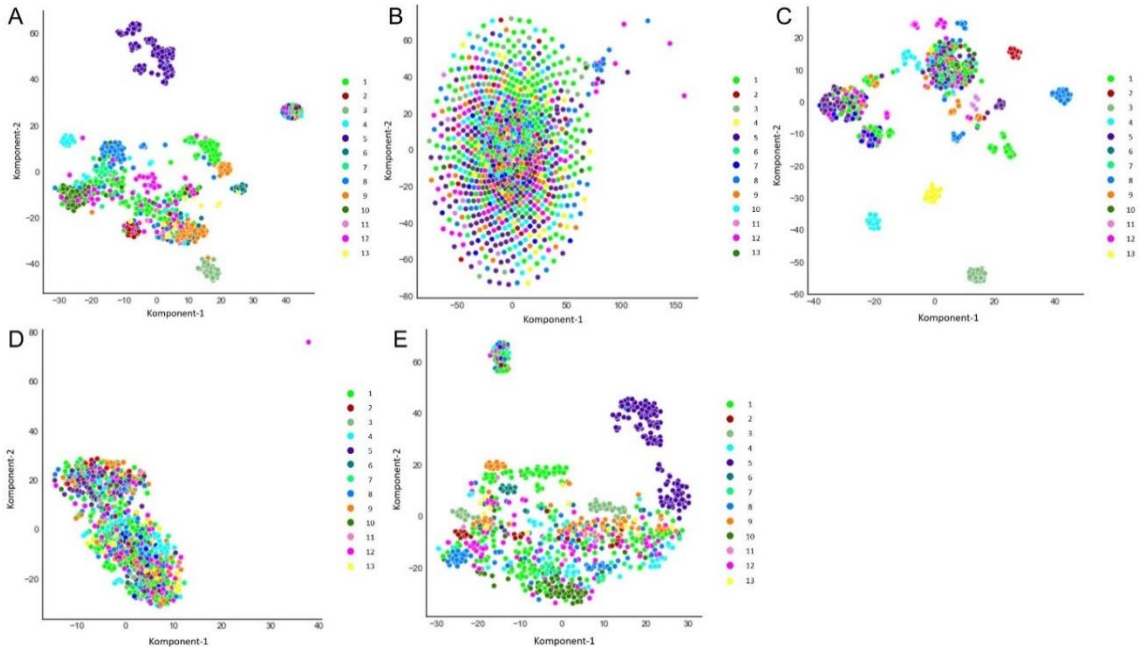
| Senaryo | İlaç sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|
| 1 | 1054 | 60 | 60678, 7 |
| 2 | 1054 | 60 | 60678, 7 |
| 3 | 1054 | 53 | 53693, 7 |
| 4 | 1054 | 7 | 6985, 7 |
| 5 | 97 | 60 | 5586, 7 |
| 6 | 957 | 60 | 55092, 7 |
| 7 | 1054 | 60 | 60678, 7 |

### 4.1.4. T-SNE Tekniğiyle Hücre Hattı Özellik Veri Tiplerinin Görselleştirilmesi

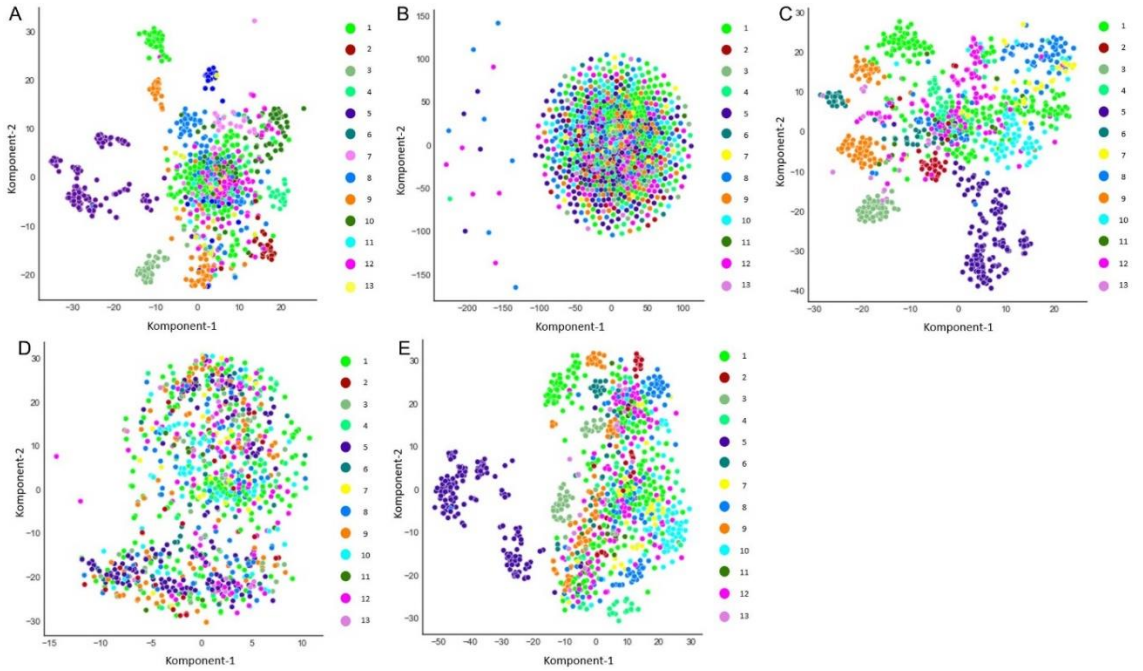
GDSC üzerinden elde edilen omik veri tiplerinin doku temelli şekilde ve 2B olarak görselleştirilmesi için t-SNE yöntemi kullanılmıştır. L1000 gen listesi filtrelemesi yapılan veri yapıları için bu yöntem yardımıyla, omik veri tiplerinin kaynak dosyaları ayrı ayrı görselleştirilmiştir (Şekil 4.6.). Farklı uzunluklara (35542 ve 3747) sahip vektörlerin alt birimleri olan ön işlemli tekil ve birleştirilmiş omik veri tiplerine ait görselleştirmeler sırasıyla Şekil 4.7. - 4.8.'de sunulmuştur. Bu iki vektöre ait şekillerin karşılaştırılmasında, gen ifade, metilasyon, birleştirilmiş omikler veri için benzer derecede dokuların ayrıldığı söylenebilir. KSD veri tipinde ise 3747 uzunluğundaki vektördeki ait grafiğin diğerine göre bir yerde kümelemediği ve bazı dokuları daha ayırt edici özellik gösterdiği görülmektedir.




**Şekil 4.6.** GDSC omik veri tiplerinin kaynak dosyaları kullanılarak t-SNE tekniğiyle iki boyutlu düzlemde görselleştirilmesi. Vektör uzunlukları L1000 gen listesi yardımıyla filtrelenmiştir. (A) gen ifade verisi (923 vektör uzunluklu); (B) mutasyon verisi (947 vektör uzunluklu); (C) metilasyon verisi (21 vektör uzunluklu); (D) KSD verisi (945 vektör uzunluklu). Şekillerdeki renkli noktalar dokulara göre şu sırayla atanmıştır; 1 (Akciğer), 2 (Böbrek), 3 (Deri), 4 (Meme), 5 (Kan), 6 (Kemik), 7 (Pankreas), 8 (Sindirim sistemi), 9 (Sinir sistemi), 10 (Solunum-sindirim sistemi), 11 (Tiroid), 12 (Ürogenital sistem), 13 (Yumuşak doku).




**Şekil 4.7.** GDSC omik veri tiplerinin ön işlemli dosyaları (toplamda 35542 vektör uzunluklu) kullanılarak t-SNE tekniğiyle iki boyutlu düzlemde görselleştirilmesi. Vektör uzunlukları L1000 gen listesi yardımıyla filtrelenmiştir. (A) gen ifade verisi (16468 vektör uzunluklu); (B) mutasyon verisi (2458 vektör uzunluklu); (C) metilasyon verisi (148 vektör uzunluklu); (D) KSD verisi (16468 vektör uzunluklu); (E) Birleştirilmiş omikler verisi (35542 vektör uzunluklu). Şekillerdeki renkli noktalar dokulara göre şu sırayla atanmıştır; 1 (Akciğer), 2 (Böbrek), 3 (Deri), 4 (Meme), 5 (Kan), 6 (Kemik), 7 (Pankreas), 8 (Sindirim sistemi), 9 (Sinir sistemi), 10 (Solunum-sindirim sistemi), 11 (Tiroid), 12 (Ürogenital sistem), 13 (Yumuşak doku).




**Şekil 4.8.** GDSC omik veri tiplerinin ön işlemli dosyaları (toplamda 3747 vektör uzunluklu) kullanılarak t-SNE tekniğiyle iki boyutlu düzlemde görselleştirilmesi. Vektör uzunlukları L1000 gen listesi yardımıyla filtrelenmiştir. (A) gen ifade verisi (897 vektör uzunluklu); (B) mutasyon verisi (963 vektör uzunluklu); (C) metilasyon verisi (955 vektör uzunluklu); (D) KSD verisi (932 vektör uzunluklu); (E) Birleştirilmiş omikler verisi (3747 vektör uzunluklu). Şekillerdeki renkli noktalar dokulara göre şu sırayla atanmıştır; 1 (Akciğer), 2 (Böbrek), 3 (Deri), 4 (Meme), 5 (Kan), 6 (Kemik), 7 (Pankreas), 8 (Sindirim sistemi), 9 (Sinir sistemi), 10 (Solunum-sindirim sistemi), 11 (Tiroid), 12 (Ürogenital sistem), 13 (Yumuşak doku).

### 4.2. Ablasyon Analizi

GDSC verisi üzerinden elde edilen her bir omik ve birleştirilmiş omikler veri tipinin ablasyon analizinde kullanılarak tahmin performansına etkisi araştırılmıştır. Oluşturulan her veri tipi için yapılan analizler aşağıda ayrı başlıklar altında verilmiştir.

#### 4.2.1. Gen İfade Verisiyle İlaç Yanıtı Tahmini

Ablasyon analizinde gen ifade verisi kullanılmasıyla DeepResponse-RF'nin gösterdiği tahmin performansları Tablo 4.18.'de verilmiştir. Tabloda, metrik temelli

olarak en yüksek tahmin performansı sunan durumlar kalın yazı tipiyle belirtilmiştir. Buna göre, çapraz doğrulamada 3. kat durumu en iyi performansı sunan durum olmuştur.

**Tablo 4.18.** GDSC gen ifade verisiyle yapılan ablasyon analizinde 5 kat çapraz doğrulama tekniğiyle elde edilen tahmin performanslarının 6 farklı skorlama metriği ile hesaplanması.

| Kat değerleri | MAE | MSE | RMSE | SCC | PCC | $R^2$ |
|---|---|---|---|---|---|---|
| 1 | 0.387 | 0.272 | 0.521 | **0.867** | 0.891 | 0.794 |
| 2 | 0.388 | 0.274 | 0.523 | 0.866 | 0.891 | 0.794 |
| 3 | **0.386** | **0.271** | **0.521** | **0.867** | **0.893** | **0.797** |
| 4 | 0.388 | 0.273 | 0.522 | 0.865 | 0.891 | 0.794 |
| 5 | 0.388 | **0.271** | **0.521** | 0.866 | 0.892 | 0.795 |

### 4.2.2. Mutasyon Verisiyle İlaç Yanıtı Tahmini

Ablasyon analizinde mutasyon verisi kullanılmasıyla DeepResponse-RF'nin gösterdiği tahmin performansları Tablo 4.19.'da verilmiştir. Tabloda, metrik temelli olarak en yüksek tahmin performansı sunan durumlar kalın yazı tipiyle belirtilmiştir. Buna göre, çapraz doğrulamada 2. kat durumu en iyi performansı sunan durum olmuştur.

**Tablo 4.19.** GDSC mutasyon verisiyle yapılan ablasyon analizinde 5 kat çapraz doğrulama tekniğiyle elde edilen tahmin performanslarının 6 farklı skorlama metriği ile hesaplanması.

| Kat değerleri | MAE | MSE | RMSE | SCC | PCC | $R^2$ |
|---|---|---|---|---|---|---|
| 1 | **0.450** | 0.364 | 0.603 | 0.821 | 0.851 | 0.724 |
| 2 | **0.450** | **0.363** | **0.602** | **0.825** | **0.852** | **0.727** |
| 3 | 0.451 | 0.366 | 0.605 | 0.822 | 0.852 | 0.725 |
| 4 | **0.450** | 0.364 | 0.604 | 0.822 | 0.851 | 0.725 |
| 5 | **0.450** | 0.364 | 0.604 | 0.822 | 0.851 | 0.724 |

### 4.2.3. Metilasyon Verisiyle İlaç Yanıtı Tahmini

Ablasyon analizinde metilasyon verisi kullanılmasıyla DeepResponse-RF'nin gösterdiği tahmin performansları Tablo 4.20.'de verilmiştir. Tabloda, metrik temelli

olarak en yüksek tahmin performansı sunan durumlar kalın yazı tipiyle belirtilmiştir. Buna göre, çapraz doğrulamada 3. kat durumu en iyi performansı sunan durum olmuştur.

**Tablo 4.20.** GDSC metilasyon verisiyle yapılan ablasyon analizinde 5 kat çapraz doğrulama tekniğiyle elde edilen tahmin performanslarının 6 farklı skorlama metriği ile hesaplanması.

| Kat değerleri | MAE | MSE | RMSE | SCC | PCC | $R^2$ |
|---|---|---|---|---|---|---|
| 1 | 0.395 | 0.281 | 0.530 | 0.862 | 0.887 | 0.786 |
| 2 | 0.395 | 0.283 | 0.532 | 0.862 | 0.887 | 0.787 |
| 3 | **0.393** | **0.280** | **0.529** | **0.863** | **0.889** | **0.790** |
| 4 | 0.394 | 0.281 | 0.531 | 0.861 | 0.887 | 0.787 |
| 5 | 0.394 | **0.280** | **0.529** | 0.862 | 0.888 | 0.788 |

### 4.2.4. KSD Verisiyle İlaç Yanıtı Tahmini

Ablasyon analizinde KSD verisi kullanılmasıyla DeepResponse-RF'nin gösterdiği tahmin performansları Tablo 4.21.'de verilmiştir. Tabloda, metrik temelli olarak en yüksek tahmin performansı sunan durumlar kalın yazı tipiyle belirtilmiştir. Buna göre, çapraz doğrulamada 3. kat durumu en iyi performansı sunan durum olmuştur.

**Tablo 4.21.** GDSC KSD verisiyle yapılan ablasyon analizinde 5 kat çapraz doğrulama tekniğiyle elde edilen tahmin performanslarının 6 farklı skorlama metriği ile hesaplanması.

| Kat değerleri | MAE | MSE | RMSE | SCC | PCC | $R^2$ |
|---|---|---|---|---|---|---|
| 1 | 0.417 | 0.315 | 0.561 | 0.845 | 0.872 | 0.761 |
| 2 | 0.417 | **0.313** | **0.560** | **0.848** | **0.874** | **0.764** |
| 3 | **0.416** | 0.314 | **0.560** | **0.848** | **0.874** | **0.764** |
| 4 | 0.418 | 0.317 | 0.563 | 0.844 | 0.872 | 0.760 |
| 5 | 0.418 | 0.314 | 0.561 | 0.846 | 0.873 | 0.762 |

### 4.2.5. Birleştirilmiş Omikler Verisiyle İlaç Yanıtı Tahmini

Ablasyon analizinde birleştirilmiş omikler verisi kullanılmasıyla DeepResponse-RF'nin gösterdiği tahmin performansları Tablo 4.22.'de verilmiştir. Tabloda, metrik temelli olarak en yüksek tahmin performansı sunan durumlar kalın yazı tipiyle

belirtilmiştir. Buna göre, çapraz doğrulamada 3. kat durumu en iyi performansı sunan durum olmuştur.

**Tablo 4.22.** GDSC birleştirilmiş omikler verisiyle yapılan ablasyon analizinde 5 kat çapraz doğrulama tekniğiyle elde edilen tahmin performanslarının 6 farklı skorlama metriği ile hesaplanması.

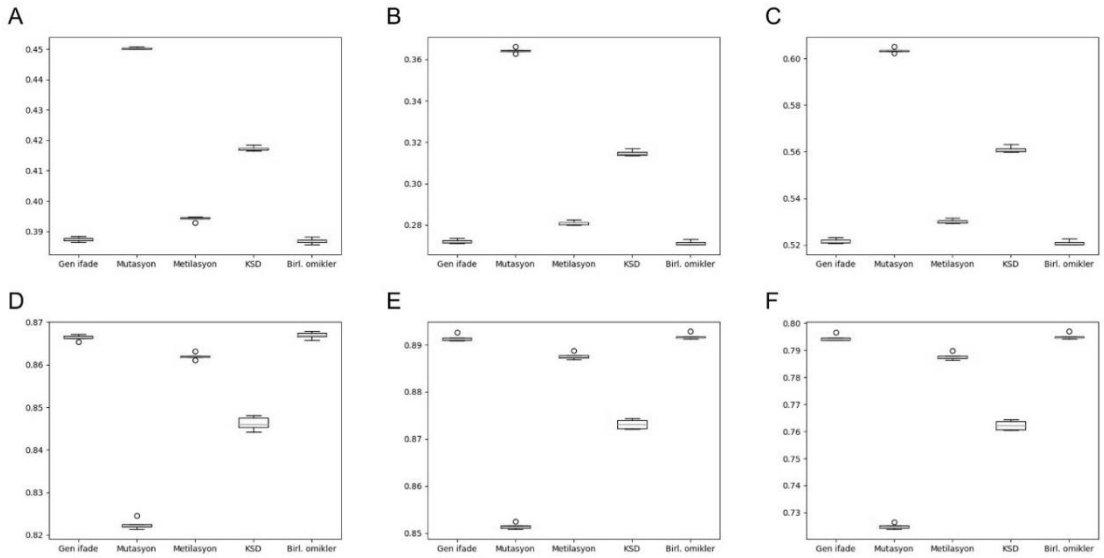
| Kat değerleri | MAE | MSE | RMSE | SCC | PCC | $R^2$ |
|---|---|---|---|---|---|---|
| 1 | **0.386** | 0.271 | **0.520** | 0.867 | 0.891 | 0.795 |
| 2 | **0.388** | 0.273 | 0.523 | 0.867 | 0.891 | 0.794 |
| 3 | **0.386** | **0.270** | **0.520** | **0.868** | **0.893** | **0.797** |
| 4 | 0.387 | 0.272 | 0.521 | 0.866 | 0.892 | 0.795 |
| 5 | 0.387 | 0.271 | **0.520** | 0.867 | 0.892 | 0.795 |

### 4.2.6. Ablasyon Analizi Sonuçlarının Kendi İçinde ve Diğer Yöntemlerle Karşılaştırılması

Ablasyon analizinde omik temelli veri tipinin kendi içlerinde karşılaştırılmıştır (Tablo 4.23.). Bu karşılaştırmanın her metrik bazında grafiksel gösterimi Şekil 4.9.'da yapılmıştır. DeepResponse-RF'nin birleştirilmiş omikler veri tipi ile kullanılmasının tüm metriklerde diğer tekil omik veri kullanımlarına performans açısından üstün geldiği görülmüştür. Diğer yandan, DeepResponse-RF ile alınan bu tahmin performanslarının RMSE, SCC ve PCC skorlama metrikleri üzerinden literatürdeki diğer yöntemlerle karşılaştırılması yapılmıştır (Tablo 4.24. – 4.26.). DeepResponse-RF'nin, yalnızca RMSE metriğinde karşılaştırılan yöntemlere üstünlük sağladığı görülmüştür.

**Tablo 4.23.** GDSC ablasyon analizinde modellenen farklı senaryolarda elde edilen sonuçların her metrik için alınan ortalamalarının karşılaştırılması.

| Yöntem adı | MAE | MSE | RMSE | SCC | PCC | $R^2$ |
|---|---|---|---|---|---|---|
| DeepResponse-RF gen ifade | **0.387** | 0.272 | 0.522 | 0.866 | 0.891 | **0.795** |
| DeepResponse-RF mutasyon | 0.450 | 0.364 | 0.604 | 0.822 | 0.851 | 0.725 |
| DeepResponse-RF metilasyon | 0.394 | 0.281 | 0.530 | 0.862 | 0.888 | 0.788 |
| DeepResponse-RF KSD | 0.417 | 0.315 | 0.561 | 0.846 | 0.873 | 0.762 |
| DeepResponse-RF birleştirilmiş omikler | **0.387** | **0.271** | **0.521** | **0.867** | **0.892** | **0.795** |




**Şekil 4.9.** GDSC ablasyon analizi senaryolarında elde edilen ilaç yanıtı tahmini performanslarının altı farklı skorlama metriği özelinde kutu grafiğiyle görselleştirilmesi. (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$.

**Tablo 4.24.** Ablasyon analizi sonuçlarının, RMSE metriği göz önüne alınarak diğer yöntemlerle karşılaştırılması.

| Yöntem adı | RMSE |
|---|---|
| DeepResponse-RF<br>birleştirilmiş omikler | **0.521** |
| DeepResponse-RF<br>gen ifade | 0.522 |
| DeepResponse-RF<br>metilasyon | 0.53 |
| DeepResponse-RF<br>KSD | 0.561 |
| DeepResponse-RF<br>mutasyon | 0.604 |
| MCA<br>gen ifade (rastgele temelli) | 1.084 |
| DualGCN<br>KSD | 1.172 |
| DualGCN<br>gen ifade | 1.191 |
| MCA<br>KSD (rastgele temelli) | 1.303 |

**Tablo 4.25.** Ablasyon analizi sonuçlarının, SCC metriği göz önüne alınarak diğer yöntemlerle karşılaştırılması.

| Yöntem adı | SCC |
|---|---|
| DeepResponse-RF birleştirilmiş omikler | 0.867 |
| DeepResponse-RF mutasyon | 0.822 |
| DeepResponse-RF KSD | 0.846 |
| DeepResponse-RF metilasyon | 0.862 |
| DeepResponse-RF gen ifade | 0.866 |
| DualGCN gen ifade | 0.887 |
| DualGCN KSD | **0.892** |

**Tablo 4.26.** Ablasyon analizi sonuçlarının, PCC metriği göz önüne alınarak diğer yöntemlerle karşılaştırılması.

| Yöntem adı | PCC |
|---|---|
| DeepResponse-RF birleştirilmiş omikler | 0.892 |
| DeepResponse-RF mutasyon | 0.851 |
| DeepResponse-RF KSD | 0.873 |
| DeepCDR gen ifade | 0.878 |
| DeepResponse-RF metilasyon | 0.888 |
| DeepCDR metilasyon | 0.89 |
| DeepResponse-RF gen ifade | 0.891 |
| DualGCN gen ifade | 0.908 |
| DualGCN KSD | **0.911** |

### 4.3. Veri İçi Alan Analizi

Veri içi alan analizinde uygulanan üç farklı bölümlendirme tipinde elde edilen sonuçlar aşağıda farklı başlıklar altında verilmiştir.

### 4.3.1. Hücre Hattı Özdeşliği Temelli Bölümlendirme (HHÖTB)

GDSC verisi üzerinden oluşturulan doku temelli veri yapıları ilk olarak HHÖTB tekniği ile analiz edilmiştir. DeepResponse-RF'nin her doku için gösterdiği tahmin performansı Tablo 4.27.'de belirtilmiştir. Sonuçlara göre, belirli bir doku üzerinden alınan tahmin performanslarının metriklerin genelinde görülecek şekilde diğerlerine üstün olmadığı görülmüştür.

EK-2'de, HHÖTB tahmin performanslarının görselleştirilmesi için dokulara özel olarak kutu grafikleri hazırlanmıştır (Şekil 8.6. – 8.18.). Bu şekillerde, her skorlama metriğine ait birer alt şekil bulunmaktadır. Alt şekillerdeki yatay eksenlerde, seçilen hiperparametrelere ait değerlerinin ikili kombinasyonları (9 adet) belirtilmiştir.

**Tablo 4.27.** GDSC'deki veri içi alan analizi sonuçlarının (HHÖTB) her dokuda metrik bazında ortalamaları alınarak sunulması.

| Doku adı | MAE | MSE | RMSE | SCC | PCC | $R^2$ |
|---|---|---|---|---|---|---|
| Akciğer | 0.533 | 0.537 | 0.701 | 0.668 | 0.711 | 0.551 |
| Böbrek | 0.514 | 0.517 | 0.689 | 0.680 | 0.723 | 0.559 |
| Deri | 0.547 | 0.557 | 0.719 | 0.666 | 0.715 | 0.530 |
| Meme | 0.528 | 0.514 | 0.693 | 0.675 | 0.721 | 0.541 |
| Kan | 0.550 | 0.554 | 0.716 | 0.667 | 0.692 | 0.527 |
| Kemik | 0.545 | 0.547 | 0.709 | **0.706** | 0.718 | 0.529 |
| Pankreas | 0.540 | 0.544 | 0.710 | 0.662 | 0.697 | 0.531 |
| Sindirim sistemi | **0.511** | **0.497** | **0.671** | 0.675 | 0.723 | **0.567** |
| Sinir sistemi | 0.537 | 0.539 | 0.697 | 0.703 | 0.723 | 0.554 |
| Solunum-sindirim sistemi | 0.525 | 0.524 | 0.689 | 0.683 | 0.726 | 0.564 |
| Tiroid | 0.542 | 0.547 | 0.704 | 0.681 | **0.740** | 0.548 |
| Ürogenital sistem | 0.578 | 0.633 | 0.759 | 0.688 | 0.712 | 0.477 |
| Yumuşak doku | 0.540 | 0.544 | 0.710 | 0.662 | 0.697 | 0.531 |
| **Ortalama değer** | **0.538** | **0.543** | **0.705** | **0.678** | **0.715** | **0.539** |

**HHÖTB Analizinden Elde Edilen Sonuçların Literatürdeki Diğer Yöntem Performanslarıyla Karşılaştırılması**

DeepResponse-RF'nin diğer yöntemlerle karşılaştırılabilmesi için HHÖTB'de elde edilen doku temelli sonuçların her metrik için ortalamaları alınmıştır (Tablo 4.28.). Yöntemimizin MSE, RMSE ve $R^2$ metriklerinde diğerlerine tahmin performansı açısından üstün geldiği görülmüştür.

**Tablo 4.28.** GDSC veri içi alan analizi (HHÖTB) sonuçlarının diğer yöntemlerle karşılaştırılması.

| Yöntem adı | MAE | MSE | RMSE | SCC | PCC | $R^2$ | Yöntemin yer aldığı çalışma |
|---|---|---|---|---|---|---|---|
| ENet | - | - | 2.216 | - | **0.448** | - | (95) |
| SRMF | - | - | 1.865 | - | 0.747 | - | |
| DrugCell | - | - | 2.392 | - | 0.616 | - | |
| tCNNs | - | - | 1.519 | - | 0.843 | - | |
| DeepCDR | - | - | 1.127 | - | 0.907 | - | |
| GraphDRP(GIN) | - | - | 1.561 | - | 0.826 | - | |
| VAE+MLP | - | - | 1.406 | - | 0.878 | - | |
| SWnet | - | 2.279 | - | - | - | 0.680 | (96) |
| DeepResponse-RF | 0.538 | **0.543** | **0.705** | 0.678 | 0.715 | **0.539** | - |

DeepResponse-RF için HHÖTB sonuçlarının ortalamaları alınmıştır.

### 4.3.2. İlaç Özdeşliği Temelli Bölümlendirme (İÖTB)

GDSC verisi üzerinden oluşturulan doku temelli veri yapıları ikinci olarak İÖTB tekniği ile analiz edilmiştir. DeepResponse-RF'nin her doku için gösterdiği tahmin performansı Tablo 4.29.'da belirtilmiştir. Sonuçlara göre, belirli bir doku üzerinden alınan tahmin performanslarının metriklerin genelinde görülecek şekilde diğerlerine üstün olmadığı görülmüştür. Ancak, hata temelli metriklerde meme dokusunda; diğer metriklerde ise ürogenital sistem dokusunda alınan tahmin performansının diğer dokulara göre daha iyi olduğu belirlenmiştir.

EK-2'de, İÖTB tahmin performanslarının görselleştirilmesi için dokulara özel olarak kutu grafikleri hazırlanmıştır (Şekil 8.19. – 8.31.). Bu şekillerde, her skorlama

metriğine ait birer alt şekil bulunmaktadır. Alt şekillerdeki yatay eksenlerde, seçilen hiperparametrelere ait değerlerinin ikili kombinasyonları (9 adet) belirtilmiştir.

**Tablo 4.29.** GDSC'deki veri içi alan analizi sonuçlarının (İÖTB) her dokuda metrik bazında ortalamaları alınarak sunulması.

| Doku adı | MAE | MSE | RMSE | SCC | PCC | $R^2$ |
|---|---|---|---|---|---|---|
| Akciğer | 0.828 | 1.150 | 1.066 | 0.302 | 0.331 | 0.020 |
| Böbrek | 0.849 | 1.230 | 1.101 | 0.249 | 0.261 | -0.077 |
| Deri | 0.822 | 1.146 | 1.062 | 0.335 | 0.327 | -0.003 |
| Meme | **0.777** | **1.025** | **1.008** | 0.340 | 0.386 | 0.075 |
| Kan | 0.817 | 1.144 | 1.060 | 0.298 | 0.325 | -0.010 |
| Kemik | 0.813 | 1.122 | 1.051 | 0.316 | 0.303 | 0.008 |
| Pankreas | 0.803 | 1.126 | 1.051 | 0.297 | 0.311 | 0.011 |
| Sindirim sistemi | 0.810 | 1.154 | 1.067 | 0.265 | 0.274 | -0.003 |
| Sinir sistemi | 0.815 | 1.142 | 1.064 | 0.291 | 0.285 | 0.012 |
| Solunum-sindirim sistemi | 0.801 | 1.091 | 1.032 | 0.288 | 0.375 | 0.094 |
| Tiroid | 0.841 | 1.198 | 1.091 | 0.306 | 0.328 | 0.019 |
| Ürogenital sistem | 0.814 | 1.110 | 1.047 | **0.377** | **0.403** | **0.110** |
| Yumuşak doku | 0.831 | 1.172 | 1.077 | 0.284 | 0.300 | -0.024 |
| **Ortalama değer** | **0.817** | **1.139** | **1.060** | **0.304** | **0.324** | **0.018** |

**İÖTB Analizinden Elde Edilen Sonuçların Literatürdeki Diğer Yöntem Performanslarıyla Karşılaştırılması**

DeepResponse-RF'nin diğer yöntemlerle karşılaştırılabilmesi için İÖTB'de elde edilen doku temelli sonuçların her metrik için ortalamaları alınmıştır (Tablo 4.30.). Karşılaştırmalar sonucunda DeepResponse-RF'nin diğer yöntemlere karşı üstünlüğü sadece RMSE metriğinde net olarak görülmektedir.

### 4.3.3. Rastgele Bölümlendirme (RB)

GDSC verisi üzerinden oluşturulan doku temelli veri yapıları üçüncü olarak RB tekniği ile analiz edilmiştir. DeepResponse-RF'nin her doku için gösterdiği tahmin performansı Tablo 4.31.'de belirtilmiştir. Sonuçlara göre, belirli bir doku üzerinden alınan tahmin performanslarının metriklerin genelinde görülecek şekilde diğerlerine üstün olmadığı görülmüştür.

**Tablo 4.30.** GDSC veri içi alan analizi (İÖTB) sonuçlarının diğer yöntemlerle karşılaştırılması.

| Yöntem adı | MAE | MSE | RMSE | SCC | PCC | $R^2$ | Yöntemin yer aldığı çalışma |
|---|---|---|---|---|---|---|---|
| SRMF | - | - | 1.828 | - | **0.654** | - | (95) |
| DrugCell | - | - | 2.388 | - | 0.362 | - | |
| tCNNs | - | - | 2.393 | - | 0.514 | - | |
| DeepCDR | - | - | 1.999 | - | 0.635 | - | |
| GraphDRP(GIN) | - | - | 2.894 | - | [-0.065, 0.423] | - | |
| VAE+MLP | - | - | 2.369 | - | 0.316 | - | |
| DeepResponse-RF | 0.817 | 1.139 | **1.060** | 0.304 | 0.324 | 0.018 | - |

DeepResponse-RF için İÖTB sonuçlarının ortalamaları alınmıştır.

EK-2'de, RB tahmin performanslarının görselleştirilmesi için dokulara özel olarak kutu grafikleri hazırlanmıştır (Şekil 8.32 – 8.44). Bu şekillerde, her skorlama metriğine ait birer alt şekil bulunmaktadır. Alt şekillerdeki yatay eksenlerde, seçilen hiperparametrelere ait değerlerinin ikili kombinasyonları (9 adet) belirtilmiştir.

**Tablo 4.31.** GDSC'deki veri içi alan analizi sonuçlarının (RB) her dokuda metrik bazında ortalamaları alınarak sunulması.

| Doku adı | MAE | MSE | RMSE | SCC | PCC | $R^2$ |
|---|---|---|---|---|---|---|
| Akciğer | 0.518 | 0.514 | 0.680 | 0.679 | 0.721 | 0.570 |
| Böbrek | **0.494** | 0.492 | 0.665 | 0.694 | 0.733 | 0.582 |
| Deri | 0.511 | 0.500 | 0.671 | 0.707 | 0.743 | 0.587 |
| Meme | 0.509 | 0.488 | 0.670 | 0.698 | 0.732 | 0.572 |
| Kan | 0.530 | 0.526 | 0.691 | 0.682 | 0.706 | 0.551 |
| Kemik | 0.517 | 0.501 | 0.673 | **0.725** | 0.730 | 0.573 |
| Pankreas | 0.521 | 0.516 | 0.685 | 0.675 | 0.709 | 0.554 |
| Sindirim sistemi | 0.498 | **0.483** | **0.656** | 0.683 | 0.728 | 0.580 |
| Sinir sistemi | 0.508 | 0.497 | 0.663 | 0.722 | 0.734 | 0.585 |
| Solunum-sindirim sistemi | 0.507 | 0.500 | 0.667 | 0.693 | 0.733 | 0.586 |
| Tiroid | 0.513 | 0.503 | 0.673 | 0.711 | **0.749** | **0.592** |
| Ürogenital sistem | 0.546 | 0.560 | 0.716 | 0.692 | 0.722 | 0.561 |
| Yumuşak doku | 0.515 | 0.509 | 0.675 | 0.688 | 0.720 | 0.572 |
| **Ortalama değer** | **0.514** | **0.507** | **0.676** | **0.696** | **0.728** | **0.574** |

**RB analizinden elde edilen sonuçların diğer bölümlendirme teknikleri ve literatürdeki yöntem performanslarıyla karşılaştırılması**

DeepResponse-RF'nin diğer yöntemlerle karşılaştırılabilmesi için RB'de elde edilen doku temelli sonuçların her metrik için ortalamaları alınmıştır (Tablo 4.32.). Yöntemimizin diğer yöntemlerle karşılaştırılması metrik temelli olarak yapılmıştır. Buna bağlı olarak, metrik temelli karşılaştırma tabloları sırasıyla şu şekildedir; MSE (Tablo 4.33.), RMSE (Tablo 4.34.), SCC (Tablo 4.35.), PCC (Tablo 4.36.), $R^2$ (Tablo 4.37.). Karşılaştırlmalarda, DeepResponse-RF'nin MSE ve RMSE metriklerinde diğer yöntemlere tahmin performansı açısından üstün geldiği görülmüştür.

**Tablo 4.32.** GDSC veri içi alan analizinde doku temelli uygulama ve farklı bölümlendirme teknikleri kullanılarak elde edilen sonuçların karşılaştırılması.

| Yöntem ve bölümlendirme tekniği | MAE | MSE | RMSE | SCC | PCC | $R^2$ |
|---|---|---|---|---|---|---|
| DeepResponse-RF (HHÖTB) | 0.538 | 0.543 | 0.705 | 0.678 | 0.715 | 0.539 |
| DeepResponse-RF (İÖTB) | 0.817 | 1.139 | 1.060 | 0.304 | 0.324 | 0.018 |
| DeepResponse-RF (RB) | **0.514** | **0.507** | **0.676** | **0.696** | **0.728** | **0.574** |

Her metrik için doku sonuçlarının ortalamaları belirtilmiştir.

**Tablo 4.33.** GDSC veri içi alan analizi sonuçlarının (RB ve MSE metriği) diğer yöntemlerle karşılaştırılması.

| Yöntem adı | MSE | Yöntemin yer aldığı çalışma |
|---|---|---|
| SWnet | 0.938 | (96) |
| WGRMF | 0.984 | |
| SRMF | 0.987 | |
| GraphDRP | 1.259 | |
| KBMTL | 1.264 | |
| CDRscan | 2.153 | |
| DeepResponse-RF | **0.507** | - |

DeepResponse-RF için doku temelli uygulama ve RB yöntemi kullanılarak MSE metriği için elde edilen sonuçların ortalamaları alınmıştır.

**Tablo 4.34.** GDSC veri içi alan analizi sonuçlarının (RB ve RMSE metriği) diğer yöntemlerle karşılaştırılması.

| Yöntem adı | RMSE | Yöntemin yer aldığı çalışma |
|---|---|---|
| RF | 2.27 | (53) |
| Rigde regression | 2.37 | |
| DeepCDR | 1.058±0.006 | |
| CDRScan | 1.982±0.005 | |
| MOLI | 2.282±0.008 | |
| SRMF (ilaç yanıtı + gen ifade) | 1.43 | (97) |
| SRMF (ilaç yanıtı) | 1.45 | |
| KBMF | 1.59 | |
| DLN | 2.08 | |
| DualGCN | 1.079 ± 0.007 | (98) |
| DeepCDR (-) | 1.265 ± 0.020 | |
| Lasso | 1.284 ± 0.007 | (98) |
| SVM | 3.115 ± 0.053 | |
| pairwiseMKL | 1.682 | (99) |
| EN | 1.839 | |
| KronRLS-MKL | 1.899 | |
| DeepCDR | 1.058 | (96) |
| GraphDRP(GIN) | 1.111 | |
| SRMF | 1.731 | |
| tCNNs | 1.782 | |
| DrugCell | 1.998 | |
| ENet | 2.368 | |
| WGRMF | 1.37 ± 0.35 | (100) |
| MCA | 0.89 | (47) |
| DeepResponse-RF | **0.676** | - |

DeepResponse-RF için doku temelli uygulama ve RB yöntemi kullanılarak RMSE metriği için elde edilen sonuçların ortalamaları alınmıştır.

**Tablo 4.35.** GDSC veri içi alan analizi sonuçlarının (RB ve SCC metriği) yöntemlerle karşılaştırılması.

| Yöntem adı | SCC | Yöntemin yer aldığı çalışma |
|---|---|---|
| DeepCDR | 0.903 ± 0.004 | (53) |
| tCNN | 0.862 ± 0.006 | |
| CDRScan | 0.852 ± 0.003 | |
| MOLI | 0.782 ± 0.005 | |
| RF | 0.767 | |
| Rigde regression | 0.731 | |
| DualGCN | **0.907 ± 0.002** | (98) |
| DeepCDR (-) | 0.877 ± 0.004 | |
| Lasso | 0.873 ± 0.002 | |

**Tablo 4.35. (Devam)** GDSC veri içi alan analizi sonuçlarının (RB ve SCC metriği) yöntemlerle karşılaştırılması.

| | | |
|---|---|---|
| SVM | 0.230 ± 0.071 | (98) |
| DeepResponse-RF | 0.728 | - |

DeepResponse-RF için doku temelli uygulama ve RB yöntemi kullanılarak SCC metriği için elde edilen sonuçların ortalamaları alınmıştır.

**Tablo 4.36.** GDSC veri içi alan analizi sonuçlarının (RB ve PCC metriği) diğer yöntemlerle karşılaştırılması.

| **Yöntem adı** | **PCC** | **Yöntemin yer aldığı çalışma** |
|---|---|---|
| DeepCDR | 0.923 ± 0.006 | (53) |
| tCNN | 0.885 ± 0.008 | |
| CDRScan | 0.871 ± 0.004 | |
| MOLI | 0.813 ± 0.007 | |
| RF | 0.81 | (53) |
| Rigde regression | 0.78 | |
| SRMF (ilaç yanıtı + gen ifade) | 0.62 | (97) |
| SRMF (ilaç yanıtı) | 0.59 | |
| KBMF | 0.49 | |
| DLN | 0.44 | |
| DualGCN | **0.925 ± 0.001** | (98) |
| DeepCDR (-) | 0.900 ± 0.004 | |
| Lasso | 0.893 ± 0.002 | |
| pairwiseMKL | 0.858 | (99) |
| KronRLS-MKL | 0.849 | |
| GraphDRP(GIN) | 0.928 | (96) |
| SRMF | 0.787 | |
| ENet | 0.78 | |
| DrugCell | 0.766 | |
| WGRMF | 0.64 ± 0.16 | (100) |
| MCA | 0.93 | (47) |
| DeepResponse-RF | 0.696 | - |

DeepResponse-RF için doku temelli uygulama ve RB yöntemi kullanılarak PCC metriği için elde edilen sonuçların ortalamaları alınmıştır.

**Tablo 4.37.** GDSC veri içi alan analizi sonuçlarının (RB ve $R^2$ metriği) diğer yöntemlerle karşılaştırılması.

| **Yöntem adı** | **$R^2$** | **Yöntemin yer aldığı çalışma** |
|---|---|---|
| MCA | 0.86 | (47) |
| Random Forest | 0.13 | (101) |
| UnoMT | 0.53 | |
| LightGBM | 0.55 | |

**Tablo 4.37. (Devam)** GDSC veri içi alan analizi sonuçlarının (RB ve $R^2$ metriği) diğer yöntemlerle karşılaştırılması.

| | | |
|---|---|---|
| DrugCell | 0.474 | (95) |
| ENet | 0.631 | |
| CDRscan | 0.698 | |
| tCNNs | 0.780 | |
| DeepCDR | 0.842 | |
| VAE+MLP | 0.856 | |
| GraphDRP | 0.823 | (96) |
| KBMTL | 0.822 | |
| SRMF | 0.861 | |
| WGRMF | 0.862 | |
| SWnet | **0.868** | |
| DeepResponse-RF | 0.574 | - |

DeepResponse-RF için doku temelli uygulama ve RB yöntemi kullanılarak $R^2$ metriği için elde edilen sonuçların ortalamaları alınmıştır.

### 4.4. Çapraz alan analizi

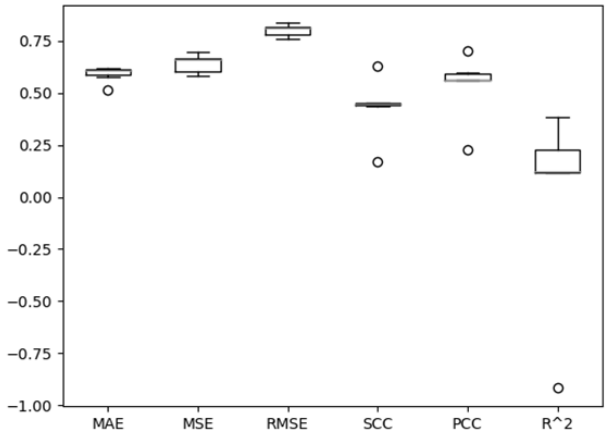
Bu analiz tipinde, DeepResponse-RF ile gerçekleştirilen GDSC – CCLE ve GDSC – NCI-60 arası analizlerden elde edilen sonuçlar aşağıda ayrı başlıklarda sunulmuştur.

#### 4.4.1. GDSC – CCLE Arası Çapraz Alan Analizi

GDSC – CCLE arası çapraz alan analizi için oluşturulan 7 senaryoda DeepResponse-RF'nin tahmin performansları Tablo 4.38.'de belirtilmiştir. Senaryolardan alınan tahmin performanslarına göre, hata temelli metriklerde 2 numaralı; diğer metriklerde ise 5 numaralı senaryo diğer durumlara üstün gelmiştir. Tahmin performanslarının görselleştirilmesi ise skorlama metrikleri temel alınarak Şekil 4.49.'da yapılmıştır.

**Tablo 4.38.** GDSC-CCLE arası çapraz alan analizlerinde uygulanan 7 senaryoda modellerin tahmin performanslarının 6 farklı skorlama metriği ile hesaplanması.

| Senaryo | MAE | MSE | RMSE | SCC | PCC | $R^2$ |
|---|---|---|---|---|---|---|
| 1 | 0.577 | **0.580** | **0.762** | 0.447 | 0.587 | 0.227 |
| 2 | **0.514** | **0.580** | **0.762** | 0.451 | 0.596 | 0.227 |
| 3 | 0.609 | 0.659 | 0.812 | 0.446 | 0.561 | 0.116 |
| 4 | 0.617 | 0.673 | 0.820 | 0.438 | 0.563 | 0.123 |
| 5 | 0.620 | 0.631 | 0.794 | **0.629** | **0.701** | **0.385** |
| 6 | 0.601 | 0.699 | 0.836 | 0.170 | 0.229 | -0.918 |
| 7 | 0.611 | 0.662 | 0.814 | 0.444 | 0.562 | 0.118 |




**Şekil 4.10.** GDSC-CCLE arası çapraz alan analizlerinde tasarlanan modellerin tahmin performanslarının değerlendirilmesi için kullanılan 6 farklı skorlama metriğiyle elde edilen sonuçların kutu grafiğiyle görselleştirilmesi.

**GDSC – CCLE Çapraz Alan Analizi Sonuçlarının Literatürdeki Diğer Yöntemlerle Karşılaştırılması**

DeepResponse-RF'nin diğer yöntemlerle (eğitim verisi olarak GDSC, test verisi olarak CCLE verisini kullananlar) karşılaştırılabilir olması için, çapraz alan analizinde iki platformda da tüm verinin kullanıldığı 7 numaralı senaryo tercih edilmiştir (Tablo 4.39.). Yalnızca $R^2$ metriği ile yapılan karşılaştırmada UnoMT, tahmin performansı açısından tüm yöntemlere üstün gelmiştir.

**Tablo 4.39.** GDSC – CCLE arası çapraz alan analizi (Senaryo 7) sonuçlarının diğer yöntemlerle karşılaştırılması.

| Yöntem adı | MAE | MSE | RMSE | SCC | PCC | $R^2$ | Yöntemin yer aldığı çalışma |
|---|---|---|---|---|---|---|---|
| Random Forest | - | - | - | - | - | 0.17 | (101) |
| LightGBM | - | - | - | - | - | 0.41 | |
| UnoMT | - | - | - | - | - | **0.50** | |
| DeepResponse-RF | 0.611 | 0.662 | 0.814 | 0.444 | 0.562 | 0.118 | - |

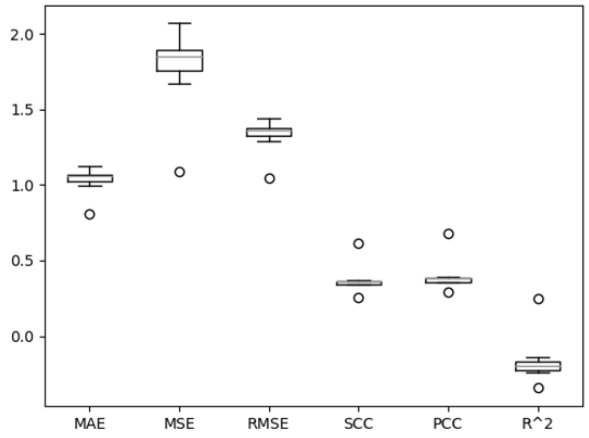
### 4.4.2. GDSC – NCI-60 Arası Çapraz Alan Analizi

GDSC – NCI-60 arası çapraz alan analizleri için oluşturulan 7 senaryoda DeepResponse-RF'nin tahmin performansları Tablo 4.40.'ta belirtilmiştir. Senaryolardan alınan tahmin performanslarına göre, tüm metriklerde 5 numaralı senaryo diğerlerine

üstün gelmiştir. Tahmin performanslarının görselleştirilmesi ise skorlama metrikleri temel alınarak Şekil 4.11.'de yapılmıştır.

**Tablo 4.40.** GDSC - NCI-60 arası çapraz alan analizlerinde uygulanan 7 senaryoda modellerin tahmin performanslarının 6 farklı skorlama metriği ile hesaplanması.

| Senaryo | MAE | MSE | RMSE | SCC | PCC | $R^2$ |
|---|---|---|---|---|---|---|
| 1 | 1.057 | 1.849 | 1.360 | 0.365 | 0.383 | -0.200 |
| 2 | 1.127 | 2.070 | 1.439 | 0.257 | 0.292 | -0.343 |
| 3 | 1.059 | 1.867 | 1.366 | 0.368 | 0.389 | -0.206 |
| 4 | 0.993 | 1.671 | 1.293 | 0.349 | 0.358 | -0.139 |
| 5 | **0.812** | **1.091** | **1.045** | **0.618** | **0.682** | **0.248** |
| 6 | 1.076 | 1.921 | 1.386 | 0.339 | 0.354 | -0.239 |
| 7 | 1.052 | 1.845 | 1.358 | 0.365 | 0.384 | -0.197 |




**Şekil 4.11.** GDSC - NCI-60 arası çapraz alan analizlerinde tasarlanan modellerin tahmin performanslarının değerlendirilmesi için kullanılan 6 farklı skorlama metriğiyle elde edilen sonuçların kutu grafiğiyle görselleştirilmesi.

**GDSC – NCI-60 Çapraz Alan Analizi Sonuçlarının Literatürdeki Diğer Yöntemlerle Karşılaştırılması**

DeepResponse-RF'nin diğer yöntemlerle (eğitim verisi olarak GDSC, test verisi olarak NCI-60 verisini kullananlar) karşılaştırılabilir olması için, çapraz alan analizinde iki platformda da tüm verinin kullanıldığı 7 numaralı senaryo tercih edilmiştir (Tablo 4.41.). Yalnızca $R^2$ metriği ile yapılan karşılaştırmada Random Forest'in, tahmin performansı açısından tüm yöntemlere üstün geldiği görülmüştür.

**Tablo 4.41.** GDSC – NCI-60 arası çapraz alan analizi (Senaryo 7) sonuçlarının diğer yöntemlerle karşılaştırılması.

| Yöntem adı | MAE | MSE | RMSE | SCC | PCC | $R^2$ | Yöntemin yer aldığı çalışma |
|---|---|---|---|---|---|---|---|
| Random Forest | - | - | - | - | - | **0.33** | (101) |
| LightGBM | - | - | - | - | - | 0.30 | |
| UnoMT | - | - | - | - | - | 0.32 | |
| DeepResponse-RF | 1.052 | 1.845 | 1.358 | 0.365 | 0.384 | -0.197 | - |

### 4.5. DeepResponse-RF ile Üretilen *In Vitro*'da Doğrulaması Yapılacak İlaç Yanıtı Tahminleri

DeepResponse-RF'nin GDSC sindirim sistemi dokusuna ait veriyle eğitilip KanSiL Lab envanterinde bulunan hücre hattı ve ilaçlardan oluşan çiftler için tahminler oluşturulmuştur. Bu tahminlere ait çıktıların, isim içeren sütunlara göre alfabetik sıralanmasıyla oluşan tablodan ilk 20 çiftin değeri Tablo 4.42.'de gösterilmiştir.

**Tablo 4.42.** DeepResponse-RF ile GDSC sindirim sistemi verisi ile eğitilip KanSiL Lab hücre hattı – ilaç çiftleri için verilen ilaç yanıtı tahminleri.

| Hücre hattı Adı | İlaç Adı | DeepResponse-RF Tahmini (pIC50) | Gerçek Değer (pIC50) | Tahmini ve Gerçek Değer Farkı |
|---|---|---|---|---|
| CAMA-1 | Camptothecin | 6.719 | 6.274 | 0.445 |
| CAMA-1 | Cisplatin | 4.239 | 4.385 | 0.146 |
| CAMA-1 | Dactolisib | 6.443 | 7.151 | 0.708 |
| CAMA-1 | Fludarabine | 3.608 | 4.089 | 0.481 |
| CAMA-1 | PI-103 | 5.21 | 5.789 | 0.579 |
| CAMA-1 | Ruxolitinib | 4.192 | 4.048 | 0.144 |
| CAMA-1 | Selisistat | 3.63 | 3.826 | 0.196 |
| CAMA-1 | Sorafenib | 4.932 | 5.191 | 0.259 |
| CAMA-1 | Staurosporine | 7.176 | 6.514 | 0.662 |
| HCT-116 | Camptothecin | 7.209 | 7.297 | 0.087 |
| HCT-116 | Cisplatin | 5.001 | 5.222 | 0.221 |
| HCT-116 | Dactolisib | 6.889 | 6.813 | 0.076 |
| HCT-116 | Fludarabine | 3.979 | 3.952 | 0.027 |
| HCT-116 | PI-103 | 5.912 | 6.232 | 0.32 |
| HCT-116 | Ruxolitinib | 4.385 | 4.355 | 0.03 |
| HCT-116 | Selisistat | 3.875 | 3.846 | 0.029 |
| HCT-116 | Sorafenib | 5.09 | 5.093 | 0.004 |
| HCT-116 | Staurosporine | 7.504 | 7.582 | 0.079 |
| Hep3B2-1-7 | Camptothecin | 6.777 | 6.681 | 0.096 |
| Hep3B2-1-7 | Cisplatin | 4.765 | 4.876 | 0.111 |

En yüksek değerli olan 20 çift için değerler gösterilmiştir.

### 4.6. DrugBank Kaynaklı İlaçlar İçin GDSC Verisi Üzerinden İlaç Yanıtı Tahminleri Üretilmesi

GDSC sindirim sistemi verisiyle eğitilen DeepResponse-RF modeliyle, aynı veride bulunan hücre hatları ve DrugBank veri tabanına ait ilaçların kombinasyonları için ilaç yanıtı tahminleri oluşturulmuştur. Bu tahminler en yüksek ilaç yanıt değerine göre sıralandıktan sonra ilk 20 hücre hattı – ilaç çiftine ait yanıtlar Tablo 4.43.'te verilmiştir.

**Tablo 4.43.** GDSC sindirim sistemi verisiyle (99 hücre hattı) ile eğitilmiş DeepResponse-RF modeli ile aynı 99 hücre hattı ve DrugBank ilaçları için oluşturulan ilaç yanıtı tahmini sonuçları.

| Hücre Hattı Adı | DrugBank İlaç Adı | DeepResponse-RF Tahmini (pIC50) |
|---|---|---|
| SNU-398 | Patupilone | 9.296232 |
| HGC-27 | Patupilone | 9.034492 |
| RKO | Patupilone | 9.028823 |
| HLE | Patupilone | 8.989822 |
| T84 | Patupilone | 8.953306 |
| HCT-15 | Patupilone | 8.882134 |
| SW620 | Patupilone | 8.873312 |
| JHH-7 | Patupilone | 8.851687 |
| HuTu-80 | Patupilone | 8.833986 |
| HuTu-80 | Elesclomol | 8.749374 |
| HCC2998 | Patupilone | 8.748809 |
| HT-29 | Patupilone | 8.736153 |
| MKN28 | Patupilone | 8.730866 |
| ETK-1 | Patupilone | 8.729611 |
| SNU-398 | Mipsagargin | 8.683044 |
| NUGC-3 | Patupilone | 8.676121 |
| SNU-423 | Patupilone | 8.672131 |
| SK-HEP-1 | Patupilone | 8.665949 |
| COLO-205 | Patupilone | 8.657949 |
| LS-180 | Patupilone | 8.646618 |

En yüksek değerli olan 20 çift için değerler gösterilmiştir.

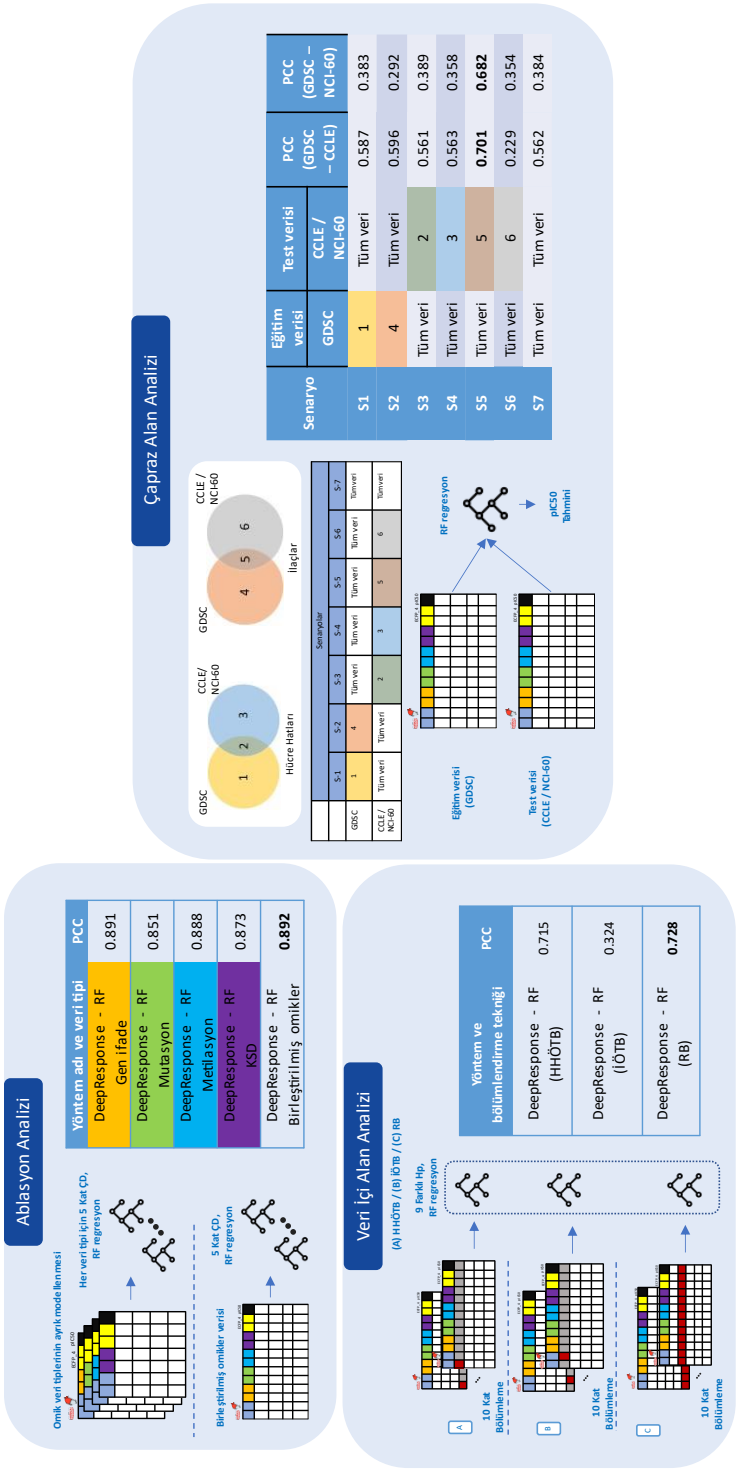
# 5. TARTIŞMA

Bu tez çalışmasıyla, ilaç yanıtı tahmini problemi için hesaplamalı bir yaklaşım sunan DeepResponse-RF yöntemi geliştirilmiştir. RF algoritmasını temel alan yöntemde, girdi verisi olarak hücre hattı özellik veri tipleri ve ilaç tanımlayıcı veri tipleri bir arada kullanılarak ilaç yanıtı tahminleri oluşturulmaktadır.

DeepResponse-RF'in alana getirdiği katkılar iki aşamada değerlendirilebilir. İlk olarak, yöntemimiz, ilaç geliştirme süreçlerinin yüksek maliyetli süreçleri ve uzun zamanda sonuç alınması gibi sorunlarını minimize edebilecek bir yaklaşım getirmektedir. Bu yaklaşımla geliştirilen ilaç yanıtı tahmini modellemesiyle, kansere karşı yüksek etkililiğe sahip ilaçların bulunmasını ve yeniden konumlandırılmasını (repurposing) destekleyici şekilde alana katkı sunmuştur. İkinci olarak ise, DeepResponse-RF'nin hücre hattı ve ilaçları temsil edici veri tiplerini beraber kullanarak ilaç yanıtı verisi olmayan hücre hattı – ilaç çiftleri için tahmin değerleri oluşturabilmesi literatüre getirilen bir diğer önemli katkı olarak sunulmuştur.

Hücre hatlarının moleküler özelliklerini temsil eden omik veri tipleri için GDSC, CCLE ve NCI-60 panellerinden yararlanılmıştır. İlaç tanımlayıcı özelliğe sahip ilaç parmak izlerinin yaratılması için ise standart SMILES dizileri kullanılmıştır. DeepResponse-RF, ilaç yanıtı tahminlerini (model çıktısı) pIC50 değeri olarak vermektedir.

Tez çalışmasının ana hedeflerine ulaşılıp ulaşılamadığının değerlendirilebilmesi için DeepResponse-RF'nin üzerinde üç farklı analiz tipi uygulanmıştır. Analiz tiplerinin genel sonuçları Şekil 5.1.'de özetlenmektedir. İlk analiz tipi olan ablasyon analizinde, dört farklı omik verinin (GDSC kaynaklı) bir arada kullanılmasının tekil kullanıma karşı tahmin performansını nasıl etkilediği incelenmiştir. Literatürdeki birçok araştırmayla örtüşecek şekilde gen ifade (Gİ) verisi en iyi tahmin performansı sergileyenlerden biri olmuştur. Genel olarak bakıldığında ise, birleştirilmiş omikler (BO) verisinin kullanımının tahmin performansını daha iyi seviyelere getirmesi skorlama metriklerinin çoğunda görülen




**Şekil 5.1.** Gerçekleştirilen analiz tiplerine ait çıktıların bir arada özetlenmesi.

değerlerle ortaya konulmuştur. Diğer yöntemlerle karşılaştırıldığında (Bkz. Tablo 4.24. – 4.26.), DeepResponse-RF-BO'un RMSE metriğinde MCA ve DualGCN; PCC metriğinde DeepCDR yöntemlerine karşı üstünlüğü görülmektedir. Bununla beraber, Gİ değerleri BO ile çok yakın olduğundan yukarıdaki karşılaştırmalarda DeepResponse-RF-Gİ'nin de benzer yöntemlerden daha iyi performans gösterdiği söylenebilir. Buradan elde edilen sonuçlar, BO'nun bilgilendirici nitelikte olmasını gösterip, sonraki aşamalarda uygulanacak analiz tiplerinde tahmin performansını iyileştireceği ihtimalini güçlendirmiştir.

Veri içi alan analiz adı verilen ikinci analiz tipinde, doku temelli yaklaşımla oluşturulan veri üzerinde katı şekilde (hücre hattı -HHÖTB- veya ilaç özdeşliği -İÖTB- üzerinden) ve rastgele (RB) şekilde bölümlendirme yapılmıştır. Modellemeler için 9 farklı RF regresyon hiperparametre kombinasyonu kullanılmıştır. Katı bölümlendirmelerdeki amaç, model eğitimi için kullanılacak veriyle test verisi arasında ortaklık (bölümlendirme türüne göre hücre hattı veya ilaç ismi) olmamasını sağlamaktır. Böylece, yeni bir hücre hattı veya ilacın göstereceği ilaç yanıtı tahmininin eğitilen model üzerinden çıktı olarak alınabilmesi kurgulanmıştır.

HHÖTB sonuçlarına göre, sindirim sistemi hücre hatlarıyla eğitilen model hata temelli metriklerde en iyi performansı göstermiştir (Bkz. Tablo 4.27.). Diğer yandan, SCC metriğinde kemik dokusu; PCC metriğinde ise tiroid dokusu en iyi tahmin performansını sunmuştur. Sinir sistemi dokusu ile alınan tahminler gerçek veriyle daha uyumlu olduğu için $R^2$ metriğinde diğer dokulara üstün gelmiştir. İÖTB ile alınan sonuçlarda, meme dokusu ile hata temelli metriklerde en iyi tahmin sonuçlarına ulaşılmıştır (Bkz. Tablo 4.29.). Korelasyon metriklerinde ise ürogenital sistem dokusu diğerlerine göre daha iyi tahmin performansı sunmuştur. RB'de ise, sindirim sistemi ve böbrek dokusu hata temelli metriklerde diğerlerine üstün gelmiştir (Bkz. Tablo 4.31.). SCC metriğinde kemik dokusu; PCC ve $R^2$ metriğinde tiroid dokusu üstün tahmin performansı göstermiştir. Sonuçların literatürdeki diğer yöntemlerle karşılaştırılabilmesi için, tüm dokularda elde

edilen performans değerlerinin kullanılan skorlama metrikleri bazında ortalaması alınarak sunulmuştur.

Veri içi alan analizinde alınan sonuçlara göre DeepResponse-RF'nin HHÖTB'de belirli metrikler bazında farklı yöntemlere üstünlüğü görülmüştür (Bkz. Tablo 4.28.). Yöntemimizin daha iyi performans gösterdiği metotlar ve metrikler aşağıda sıralanmıştır. SWnet (MSE metriğinde); ENet, SRMF, DrugCell, tCNNs, DeepCDR, GraphDRP(GIN), VAE+MLP (RMSE metriğinde); ENet, DrugCell (PCC metriğinde).

İÖTB'de sonuçlarının karşılaştırılmasında DeepResponse-RF'nin üstün olduğu yöntemler ve metrikler şu şekildedir (Bkz. Tablo 4.30.). SRMF, DrugCell, tCNNs, DeepCDR, GraphDRP(GIN), VAE+MLP (RMSE metriğinde); GraphDRP(GIN) ve VAE+MLP (PCC metriğinde).

Üçüncü bölümlendirme yöntemi olan RB'de ise DeepResponse-RF'nin tahmin performansının daha iyi olduğu yöntemler ve metrikler aşağıda belirtilmiştir (Bkz. Tablo 4.33 – 4.37.). SWnet, WGRMF, SRMF, GraphDRP, KBMTL, CDRscan (MSE metriğinde); MCA, DeepCDR, GraphDRP(GIN), SRMF (ilaç yanıtı + gen ifade), SRMF (ilaç yanıtı), KBMF, pairwiseMKL, SRMF, tCNNs, EN, KronRLS-MKL, DrugCell, DLN, RF, ENet, Rigde regression, DeepCDR, DualGCN, DeepCDR (-), Lasso, WGRMF, CDRScan, MOLI, SVM (RMSE metriğinde); SRMF (ilaç yanıtı + gen ifade), SRMF (ilaç yanıtı), KBMF, DLN (PCC metriğinde); LightGBM, UnoMT, DrugCell, Random Forest ($R^2$ metriğinde).

GDSC verisi üzerinde benzer bölümlendirme tekniklerini uygulayan yöntemlere karşı DeepResponse-RF'nin tahmin performansı açısından daha iyi olmasının ana nedeni olarak birleştirilmiş omikler verisinin ve ilaç tanımlayıcılarının kullanılma stratejisinin benimsenmesi gösterilebilir. Diğer yandan, DeepResponse-RF'nin düşük performans gösterdiği durumlar göz önüne alındığında ise bazı yetersizliklerin buna ön ayak olduğu söylenebilir. Veri içi alan analizinde kullanılması planlanan hücre hattı – ilaç çiftlerine ait verinin analiz edilemeyecek kadar büyük hacimli olması en büyük sorun olarak karşımıza

çıkmıştır. Sorunun çözümü için uygulanan doku temelli yaklaşımla, veri kolayca işlenebilir, depolanabilir ve analiz edilebilir hale getirilmiştir. İzlenen bu yöntemle elde edilen tahmin performansları, daha büyük yapıdaki verinin (güçlü hesaplama kapasitesine sahip sunucu veya bilgisayarlar yardımıyla) işlenmesi durumunda daha yüksek seviyelere çekilebilir. RB tahmin performansının diğer iki bölümlendirme tekniği performanslarına üstünlüğü de bu düşünceyi destekler niteliktedir. Ek olarak, karşılaştırma yaptığımız yöntemlerde de doku temelli yaklaşımın tercih edilmediği görülmüştür. Bu durum da, daha büyük veriyle eğitilen diğer yöntem modellerinin yüksek performans sergilemelerine katkı sağlayan önemli bir faktör olarak görülebilir.

En iyi sonuçların elde edildiği RB'nin ardından HHÖTB'nin gelmesi, belirlenen hücre hattı özellik vektöründe sadece hücreyi en iyi temsil edecek genlerin her omik veri tipi için kullanılmasına bağlanabilir. Bu konuda, L1000 işaretçi genlerinin filtrelenerek kullanılması asıl önemli değişkenlerin belirtilmesinde avantajı sağlayan değişiklik olmuştur. İÖTB tekniğiyle kurgulanan modellerin HHÖTB ve RB'dekilerden daha düşük performans gösterdiği skorlama metrikleri sonuçlarıyla net olarak görülmektedir. Bu konuda, oluşturulan modelin hücre hattı özelliklerinin daha önemli olduğunu tespit edemediği yorumu yapılabilir. Ek olarak, kullanılan ilaç parmak izlerinin büyük moleküllü ilaçların ayırt edici özellikleri için yeterince iyi temsil oluşturamadığı düşünülmektedir. Yöntemimizde yararlandığımız ECFP_4 tipindeki parmak izleri küçük ilaç molekülleri için avantaj sağlasa da, büyük moleküller için aynı performansı sağlayamadığı bilinmektedir. Buradan hareketle, ilaç parmak izi oluşturmada 3 adımlı yöntem (ECFP_6) izlenerek büyük moleküllerin özellikleri daha kapsamlı şekilde tespit edilebilecektir. Farklı bir yaklaşımla, isimler yerine ilaç benzerliklerine göre bölümlendirme yapılarak tahmin performansında iyileşme sağlanabileceği düşüncesindeyiz.

Gerçekleştirilen üçüncü analiz türü olan çapraz alan analizinde, GDSC verisi ile eğitilen model, farklı işlem akışlarında olacak şekilde CCLE ve NCI-60 verisiyle test edilmiştir. Tahmin performansının ölçülmesi için, hücre hattı ve ilaç ortaklıkları ve tüm

veri kullanımı üzerinden yedi farklı senaryo kurgulanmıştır (Bkz. Tablo 3.37.). Eğitim verisinde sadece iki senaryo test verisindeki ortak olmayan hücre hatları ve ilaçları içerirken; test verisinin çeşitlenmesinde hücre hatları ve ilaçlar için ayrı ayrı ortak olan / olmayan şeklinde ana verinin alt örnekleri oluşturulmuştur.

GDSC – CCLE çapraz alan analizinde senaryolar (1-7 arası numaralarla isimlendirilmiştir) gösterdikleri tahmin performansları kapsamında karşılaştırılmıştır (Bkz. Tablo 4.38.). 1 ve 4 numaralı senaryolarda kıstas olarak ortak olmayan hücre hatları alınmıştır. 1 numaranın diğerine göre tüm metriklerde üstün geldiği görülmektedir. Burada, 1 numaranın ortak olmayan GDSC hücre hatlarıyla eğitilmesi belirleyici olmuştur ve tüm CCLE hücre hattı – ilaç çiftleri için daha iyi tahmin performansı geliştirilmiştir.

2 ve 6 numaralı senaryolarda kıstas olarak ortak olmayan ilaçlar alınmıştır. Burada ise, 2 numarada modelin GDSC'deki ortak olmayan ilaçlara ait veriyle eğitilmesi diğer modele nazaran çok daha iyi düzeyde tahmin oluşturma yeteneği sağlamıştır. Sonuç olarak, 1-4 ve 2-6 arasındaki karşılaştırmalardaki benzerlikler, modelin eğitiminde kullanılan veride test verisinde bulunmayan hücre hattı veya ilaçların kullanılmasından ileri gelmiştir.

3, 5 ve 7 numaralı senaryolardaki ortaklıklara (hücre hattı veya ilaç) ve tüm veri kullanımına dayalı olan modeller karşılaştırılmıştır. Model eğitiminde GDSC'deki tüm verinin kullanıldığı bu senaryolarda test verisi olarak, ortak hücre hatlarına ait CCLE verisi (3 ve 5 için) ve tüm CCLE verisi (7 için) kullanılmıştır. Skorlama metrikleri göz önüne alındığında, 3 ve 7 numaranın birbirlerine oldukça yakın olduğu; 5 numaranın ise tüm modellerden daha üstün tahmin performansı sergilediği görülmektedir. Alınan sonuçlara bağlı olarak, eğitilen modelde de bulunan ilaçlar için verilen tahminlerin daha az hata ve yüksek korelasyon seviyelerini getirdiği saptanmıştır. Aynı durumun hücre hattı ortaklığı için geçerli olduğu görünmemektedir. Tespit edilen bu farklılığın nedenleri NCI-60 sonuçlarıyla beraber aşağıda belirtilmiştir.

GDSC – NCI-60 çapraz alan analizinde senaryolar gösterdikleri tahmin performansları kapsamında karşılaştırılmıştır (Bkz. Tablo 4.40.). 1 ve 4 numaralı senaryolarda kıstas olarak ortak olmayan hücre hatları alınmıştır. Korelasyon metriklerinde (SCC, PCC) 1'in 4'e üstün olduğu; hata temelli metriklerde (MAE, MSE, RMSE) ise tam tersi bir durum görülmektedir. Bu durumda, korelasyon metriği sonuçları iki değişken (gerçek ve tahmini değer) arası ilişkiyi daha doğru tespit edebildiği için hata temelli olanlara nazaran daha güvenilir olarak değerlendirilebilir. Bundan dolayı, 1'de modelin sadece GDSC'de bulunan hücre hatlarına ait veriyle eğitilmesiyle NCI-60 tüm verisi için verilen tahmin performansının daha iyi olduğu sonucuna varılabilir.

2 ve 6 numaralı senaryolarda kıstas olarak ortak olmayan ilaçlar alınmıştır. 6'da, modelin GDSC'nin tüm verisi kullanılarak eğitilerek aynı modele tanıtılmayan ilaçlara sahip test verisi için verdiği tahmin performansları tüm metriklerde 2'den daha yüksektir. Alınan bu sonuçların GDSC – CCLE analizinde yer alan 2-6 karşılaştırmasındaki performans seviyesini yakalayamamasının nedeni olarak 2'deki modelin daha küçük hacimli veriyle (GDSC-CCLE'de 391, GDSC-NCI-60'ta 307 ilaca ait veri) eğitilmesi gösterilebilir.

GDSC - NCI-60 analizinde 3, 5 ve 7 numaralı senaryolar, ortaklıklara (hücre hattı veya ilaç) ve tüm veri kullanımına dayalı olan modeller olduğu için karşılaştırılmıştır. Model eğitiminde GDSC'deki tüm verinin kullanıldığı bu senaryolarda test verisi olarak, ortak hücre hatlarına ait (3 için), ortak ilaçlara ait (5 için) NCI-60 verisi ve tüm NCI-60 verisi (7 için) kullanılmıştır. Hata temelli metrikler göz önüne alındığında, 7 numaranın hata metriklerinde 3 numaraya üstün olduğu; korelasyon metriklerinde ise tam tersi bir durum gözlenmiştir. Daha doğru çıkarım yapılabilmesi için korelasyon metrikleri üzerinden bakıldığında, 3'te test verisinin yalnızca panellerde ortak olan hücre hatları üzerinden oluşturulmasıyla 7'ye göre daha iyi tahmin performansı sağladığı görülmektedir. 5 ise tüm metriklerde bütün senaryolardan üstün tahmin performansı sunmuştur. Metrik sonuçları yorumlandığında, modele tanıtılmış olan ilaçlara verilen

tahminlerin daha doğru şekilde yapılabildiği görülmüştür. Ancak, benzer durumun hücre hatları için geçerli olmadığı 3'ün tahmin performansında tespit edilmiştir.

Yukarıdaki karşılaştırmalara ek olarak, GDSC – CCLE ve GDSC – NCI-60 arası analizlerin genel çıktılarında genel olarak benzer şekilde iki durum görülmektedir. Birincisi, $R^2$ metriğinde GDSC - CCLE analizindeki sonuçların oldukça düşük; GDSC - NCI-60 analizinde ise çoğu değerin negatif olması durumu. İkincisi ise, özellikle 3 ve 5 numaralı senaryoların karşılaştırılmasıyla ortaya çıkan ortak hücre hatları için tahmin performanslarının düşük düzeyde kalması durumu. Bu iki durum esasen, literatürde de genişçe işlenmiş olan standardizasyon problemine işaret etmektedir.

Çapraz alan analizlerindeki 7 numaralı senaryoların $R^2$ metrik değerleri Xia ve ark. (101)'nın makalesinde yer alan Random Forest, LightGBM, UnoMT yöntemleriyle karşılaştırılmıştır. GDSC-CCLE (Bkz. Tablo 3.39.) ve GDSC – NCI-60 (Bkz. Tablo 3.41.) sonuçlarının bahsedilen tüm yöntemler için geride kaldığı görülmüştür. Çapraz alan analizlerinde standardizasyon probleminden doğan nedenlerle negatif (sonuçların belli bir trendi takip etmemesi durumu) veya çok düşük seviyede kalan $R^2$ değerlerinin bu duruma yol açtığı söylenebilir.

DeepResponse-RF ile *in vitro* doğrulaması yapılacak hücre hattı – ilaç çiftleri, GDSC'de veri noktası bulunmayan 320 çift içinden seçilmiştir (Bkz. Tablo 4.42.). Seçilen yüksek potansiyeldeki ilaçlardan olan eprinomectin için Huh7, Hep3B, SNU 387 / 423 / 475 gibi HSK hücre hatları üzerinde *in vitro* denemeleri TÜSEB projesindeki ortağımız ODTÜ KanSiL Lab tarafından yapılmıştır. KanSiL Lab'ın gerçekleştirdiği SRB kolorimetrik analizi, RT-CES, hücre döngüsü ve Western blot analizleri sonuçlarının DeepResponse-RF ile verilen tahminler ile örtüştüğü görülmüştür. Deneylerde, eprinomectin molekülünün gösterdiği potansiyelin daha önce HSK hücre hatları için onaylanmış bir ilaç olan sorafenib ile karşılaştırılabilecek kadar iyi olduğu belirlenmiştir. Elde edilen bulgular, tasarlanan

DeepResponse-RF yönteminin ilaç yeniden konumlandırma amacıyla da kullanılabileceğini göstermiştir.

Yapılan son analizde, DrugBank kaynaklı ilaçlar için GDSC sindirim sistemi verisinde bulunan hücre hatlarına yönelik tahminler üretilmiştir (Bkz. Tablo 4.43). GDSC sindirim sistemi ile eğitilen DeepResponse-RF ile verilen bu tahmin sonuçları ilaç yantıları hakkında fikir vermesi açısından önem taşımaktadır. İleride, bu sonuçlar göz önünde bulundurularak seçilen yüksek potansiyelli ilaçlar için de *in vitro* doğrulamaları yapılabilir.

Çalışmanın ilerlemesiyle, tercih edilen veri kaynaklarının kullandığı araçlara veya yöntemlere bağlı olarak bazı konularda yetersizlikler tespit edilmiştir. Doğru ilaç yanıtının tespitine veya yüksek düzey tahmin performansının elde edilmesine etki edebileceği düşünülen önemli olan bu noktalar aşağıda belirtilmiştir.

İlaçların deney ortamında hastalar üzerinde denenmesi beklenmeyen yan etki (advers) durumları ve etik açıdan sorunlar doğurması nedeniyle ilaç yanıtı için üretilen geniş kapsamlı hasta verisi (bugünkü bilgimiz dahilinde) neredeyse hiç bulunmamaktadır. Hücre hatları ise bu duruma alternatif olarak, hastada gelişen tümörün moleküler özelliklerini modellemede ve ilaç geliştirme aşamalarında yaygın olarak kullanılmaktadır (38). Hücre hatlarının kullanımının genellikle maliyet ve hızlı sonuç alma açısından avantajları bulunsa da teknik açıdan bazı noksanlıkları bulunmaktadır. Kanser hücre hatlarının primer kültürlerden elde edilmesi sonrasında uygulanan pasajlama işlemleri ile orijinal tümörde bulunmayan varyasyonlara (örneğin, ilaç yanıtını etkileyecek genetik polimorfizmler) yol açabilmektedir (28). Diğer yandan, çevresel faktörler, tümör mikroçevresi gibi tümör davranışına etki eden unsurların iki boyutlu ortamda büyütülen hücreler için geçerli olmaması ve bu hücrelerin sitotoksik moleküllere karşı daha hassas olmaları da göz önüne alınmalıdır (18,38,102). Bahsedilen bu olgular, sonuçların ilaç yanıtı eğrisi üzerinde değerlendirilmesinde yanıltıcı olabilecek durumları oluşturabilmektedir.

Hücre hattı kültürlerinin homojen halde bulunmasını ve yukarıda bahsedilen problemlerini farklı açılardan aşmaya çalışan hasta-kaynaklı organoid (patient-derived organoids, PDO) veya ksenograft (patient-derived xenografts, PDX) kültür teknikleriyle üretilmiş veri tipleri de son yıllarda alandaki araştırmalarda yer bulmaya başlamıştır. Orijinal tümörle direkt karşılaştırılabilen bu iki teknik, özelleştirilmiş üç boyutlu kültür ortamları, benzer tümör mikroçevresi ve gibi avantajlarının olması tümör özelliklerinin daha iyi taklit edilmesine olanak sağlamaktadır (3,103). Ancak, henüz bu tekniklerin gerektirdiği prosedürler oldukça maliyetli ve zaman alıcı olduğundan özelleşmiş laboratuvarlar dışında yürütülen çalışmalar oldukça zorlayıcı olacaktır. İlaç yanıtı tahminini başarılı şekilde gerçekleştiren bir PDO modellemesi başarılı bir örnek olarak Pasch ve ark. (104)'nın kolorektal kanser için yaptığı çalışma gösterilebilir. PDX veri tiplerinin ise ilaç yanıtı modellemesi yapan araştırmalarda (52) performans doğrulama adımları için kullanıldığı belirtilmiştir.

Gelecekte, PDO ve PDX gibi tekniklerin iki boyutlu kültür tekniğine göre daha fazla kabul görüp yaygınlaşmasıyla deneysel olarak elde edilen ilaç yanıtı ilişkilerinin hasta tedavisinde yön göstericiler olarak yer edineceğini düşünmekteyiz. DeepResponse yönteminin geliştirilebilmesi için sonraki çalışmalarda performans doğrulama adımlarında üç boyutlu tekniklere ait veri tiplerinin de değerlendirilme kapsamına alınması tahminlerin güvenirliğini desteklemesi açısından önemli olabilir.

Standardizasyon problemi analizlerimizi etkileyen sorunların başında gelmektedir. Bu problem, panellerde hücre hattı özellik veri tiplerini ve ilaç yanıtlarını değerlerini oluşturmada kullanılan ekipmanların, uygulanan prosedürlerin farklılığından dolayı, belirli bir standart çıktı oluşturma yolu izlenmemesinden ortaya çıkmaktadır. Özellikle, çapraz alan analizi senaryolarındaki modellerin eğitiminde ve testinde kullanılan veri tiplerinin farklı hücre hattı panellerine ait olması nedeniyle tahmin performansları etkilenmiştir. Farklı panellere ait aynı omik veri tipleri benzer dağılımlar gösterse de, veri yapıları standart yöntemle elde edilmedikleri için yapılan ilaç yanıtı

tahminleri üzerinde olumsuz etki yaratmıştır. Ayrıca, eşleşmeyen hücre hattı ve ilaç sayılarının yüksek olması nedeniyle model eğitimi için kullanılan veri noktası sayısı oldukça azalmıştır. İlaç parmak izlerini aynı yöntemle elde edilen sayı dizisiyle tanıyan modeller (her iki analizde de 5 nolu senaryolar) bu belirtilen nedenlerle diğerlerine performans açısından üstün gelebilmiştir.

GDSC – CCLE tahmin performanslarının GDSC – NCI-60 arasındakilerden genel olarak daha iyi olduğu saptanmıştır (Bkz. Tablo 4.38. ve 4.40.). Bunun nedeni olarak ise GDSC – CCLE arası uyuşmazlıkların daha az olması gösterilebilir. Ayrıca, belirtilen iki panel arası ilişkilerin incelendiği Goodspeed ve ark. (17)'nın araştırmasıyla da örtüşecek şekilde olduğu görülmüştür. GDSC – NCI-60 arası analizde ise hem ilaç yanıtı analizi türünün ve sürelerinin hem de kullanılan ilaç yanıtı metriklerinin farklı olmasından doğan uyuşmazlıkların tahmin sonuçlarına da yansıdığını düşünmekteyiz.

Bahsedilen panellerdeki veri yapıları arası uyumsuzlukların veri içeriği bakımından en azından benzer birimlerle, aynı formatla ve ortak kullanılan isimlerle (genler, ilaçlar, hücre hatları için) beraber oluşturulması önem taşımaktadır. Veri ön işleme adımlarını kolaylaştırıcı bu düzenlemeler, tez çalışması kapsamında da yararlanılan CellMinerCDB (55,66) gibi farklı panellere ait veri tiplerini bünyesinde barındıran veri tabanları tarafından yapılmaktadır. Ön işleme adımlarında bazı ilaç yanıtı veri noktalarının ve moleküler veri tiplerinde orijinal veriye kıyasla eksikliklerin bulunması, analizler için önceliği orijinal veri yapılarına vermemize neden olmuştur.

Yukarıda bahsedilen önemli noktalar, bizlere DeepResponse-RF yönteminin hala geliştirilebilecek yönleri olduğunu göstermektedir. Bozucu faktörlerin etkisinin ortaya çıkarılması ve modelde kurulan ilişkilerin daha sağlıklı şekilde oluşturulması için yeni değişkenlerin ve daha büyük veri yapılarının analizine ihtiyaç duyulmaktadır. Böylece, doğru belirleyiciler kullanılarak tahmin performansı iyileştirilecek ve ilaç yanıtının

moleküler temellerine ışık tutulmuş olacaktır. Yöntemimizin kısıtlı tahmin kabiliyetini artırabilmek için gelecekte aşağıda belirtilen veri tiplerinden yararlanılabilir.

ECFP_4 parmak izi oluşturma tekniğinin büyük moleküllü ilaçların özelliklerini ayrıntılı şekilde çıkaramamasının tahmin performansını etkileyebileceği açıktır. İleride, bu kısıtlamayı aşabilen farklı bir teknik yönteme dahil edilebilir. Yönteme entegre edilebilecek tekniklerden biri olarak Capecchi ve ark. (76) tarafından geliştirilen MAP4 (MinHashed Atom-Pair Fingerprint) yöntemi gösterilebilir. Bu tekniğe ait araştırma makalesinde, atom çifti (*atom pair*) özelliklerini temel alarak, hem küçük hem büyük boyutlu moleküllerde diğer parmak izi çıkarma yöntemlerine göre daha iyi performans gösterdiği belirtilmiştir.

Hücrelerin ilaçlara yanıtlarında ilk olarak etkisi bulunan yapılar proteinlerdir. İlaç hedefleri, ilaç taşıyıcıları, ilaç metabolize edici enzimler gibi hücresel moleküllerin protein yapısında olması, proteinlerin hangi açılardan ilaç etkisini değiştirebileceği hakkında fikir verebilir. Hücrenin genetik yapısında meydana gelecek olan varyasyonların direkt olarak proteinler üzerinde de etkili olduğu düşünüldüğünde, proteomik gibi alt akım (downstream) moleküler veri tiplerinin de önemli olduğu açıktır. Çalışmada kullandığımız dört omik veri tipi üst akım (upstream) profillere ait olduğundan, proteomik verinin hücreyi moleküler açıdan temsilde tamamlayıcı rol oynayacağını düşünmekteyiz. Tüm panellerin proteomik veriyi sağlayamadığı bilindiğinden, bazı durumlarda gen – protein ağlarından yararlanılabilir.

Yöntemimizin genişletilmesinde yer alabilecek diğer veri tipi ise sinyal yolağı profilleridir (SYP). SYP, bir hücre hattından elde edilen ilaç yanıtında hangi yolakların etkili olduğunun yorumlanabilmesinde bilgilendirici olacaktır. Özellikle, ilaçlara karşı dirençli olan hücre hatlarının detaylı incelenip doğru yaklaşımın bulunması için bu veri tipi büyük önem arz etmektedir. Yöntemimizin devam çalışmalarında, verideki ilaçların etki ettiği

bilinen yolaklar göz önüne alınarak hangi şekilde ilaç yanıtını oluşturduğu nitel olarak araştırılabilir.

Kodlayıcı olmayan RNA'ların bazılarının ilaç yanıtı için biyobelirteç özelliğinde olduğu bildirilmektedir. Bu kapsama giren mikro RNA'lara (miRNA) ait ifade düzeyleri, dokular arasında farklılık gösterdiği, hastalık durumunda veya ilaç etkisiyle değişebildiği için dikkate değer bir faktör olarak göz önüne alınmalıdır (2,42).

İlaç yanıtı tahminlerinin doğrulanabilmesinde hastalara ait veri tiplerinin yönteme dahil edilmesi, tahminlerin gerçek duruma ne düzeyde transfer edilebileceği hakkında yorum yapma fırsatı verecektir. TCGA gibi veri tabanlarında hastaların isimsizleştirilmiş halde bulunan veri tiplerinin bu bağlamda kullanılabileceğini düşünmekteyiz. Böylelikle, hastaya özgü tedavi seçeneklerinin belirlenebilmesine yardımcı olacak modelleme senaryoları oluşturulabilecektir.

İlaç yanıtı tahmini probleminde aydınlatılmayı bekleyen birçok nokta bulunmaktadır. Gelecekte, bu araştırma alanında göz önünde bulundurulması büyük kazanımlar sağlayacak meselelere aşağıda değinilmiştir.

İlaç yanıtı tahmini yöntemlerinin nihai amacı hastaların farklı özelliklerine yönelik çözümler bulmaya yardımcı olmaktır. Bununla beraber, yöntemlere ait modellemelerin sonuçlarının kliniğe aktarılabilirliği en büyük sorun olarak karşımıza çıkmaktadır. Literatürde PDX gibi veri tipleriyle doğrulama yapan yöntemler olsa da hastanın yaşı, cinsiyeti, kanser evresi, tümör derecelendirmesi, etnisite, vücut kitle endeksi gibi kayıtların da değerlendirmeye alındığı bir yönteme rastlanmamıştır. Bu açıdan, anonim hale getirilmiş elektronik sağlık kayıtları hastaların özelliklerine göre sınıflandırılmasında önemli rol oynayacak ve bu özelliklere uyumlu ilaçların seçilebilmesine yardımcı olabilir.

Çalışmamızda, mutasyon verisi gen temelli olarak "var" ("1" atanan değeriyle) veya "yok" ("0" atanan değeriyle) şeklinde düzenlenmiştir. Bu durum, genin hangi

mutasyonun meydana geldiğini belirtmekte yetersiz kalmaktadır. Klinik değerlendirmede, ilaçların bazıları gendeki belirli mutasyonların varlığı göz önüne alınarak hastanın tedavisinde kullanılmaktadır. Buna örnek olarak, küçük hücreli olmayan akciğer kanserinde L858R (ekzon 21) değişimi için Erlotinib ve Gefitinib ilaçlarının kullanımı gösterilebilir (105). Bahsedilen bu duruma bağlı olarak, gendeki değişikliklerin açık olarak belirtilerek mutasyon verisine eklenmesi yararlı olacaktır. Böylece, bazı kanser türleri için ilaç ve kanser ilişkileri daha sağlıklı şekilde yapılabilir.

Hücre hattı panellerinde tümörlü dokuların temsillerinin artırılması gerekmektedir. Çalışmamızda doku temelli analizler için yapılan düzenlemeler sırasında bazı dokularda oldukça az sayıda hücre hattının bulunduğu görülmüştür. Panellerde incelenen dokulara ait hücre hattı sayılarının artması hem daha iyi tümör temsili sağlayacaktır hem de ilaç yanıtı düzeylerini doğru seviyede yakalayabilmemiz için fırsat verebilir.

Hücre hattı panellerinde karşılaşılan bir diğer sorun ise deneysel tekrarlardır. Deneysel tekrarlar analiz başına düşen maliyeti artırdığından bazı prosedürlere eklenmemektedir. Buna bağlı olarak, ilaç yanıtı değerlerinin doğruluğu ve kesinliği önemli ölçüde düşmektedir. Deneysel tekrarların panellerde artırılmasıyla maliyetler artsa da daha güvenilir ve veri gürültüsü daha az olan sonuçlar elde edilebilecektir.

DeepResponse-RF ve incelenen diğer yöntemler, hücre hatları üzerinde tekil olarak (*monotherapy*) denenen ilaçların etkilerini modellemektedir. Ancak, lösemi gibi bazı kanser türlerinde birden fazla ilacın beraber kullanılmasının yarattığı olumlu etki tekil kullanıma göre daha fazla olmuştur (13). Diğer yandan, GDSC gibi veri tabanlarının veri setlerini ilaç kombinasyonuna yönelik olacak şekilde genişletmektedir. İleride geliştirilecek tahmin modellerinde uygulanacak stratejilerin de bu duruma uygun şekilde oluşturulması gerekecektir.

Tez kapsamında hedeflediğimiz amaçlara ulaşabilmek için gerçekleştirilen analizlerden çok, çeşitli veri tabanlarına ait veri yapılarının düzenlenmesi aşamaları kendi açımdan daha öğretici olduğu kanaatindeyim. Kullandığımız veri yapılarının modelleme öncesi özellik seçimi aşamalarında tanınması mümkün oldu. Böylece, hücre hatlarını farklı vektör boyutlarında temsil edecek matrislerin düzenlenmesi kolaylaşmıştır. İlaç tanımlayıcı veri tiplerinin oluşturulmasında pek çok yöntem ve Python kütüphanelerinden yararlanarak eldeki bilgi yığını ile nasıl baş edebileceğimiz hakkında çok değerli tecrübeler edinmiş oldum. İlk defa, bu kadar yüksek boyutlu veri yapılarını sunucular yardımı ile analiz etme fırsatına eriştim. Zaman ve iş yükünün azaltılması açısından öncelikle küçük veri yapıları üzerinden çalışma alışkanlığını edinmenin büyük yararını görmüş durumdayım. Bir parçası olduğum TÜSEB projesi için gereken iş paketlerinin zamanında yetiştirilmesi ve raporlanması zaman kullanımı hakkında yeni alışkanlıklar geliştirmeme yardımcı olmuştur. Ve son olarak, kısa süre önce başlamış olduğum programlama, veri analizi ve makine öğrenmesi algoritmaları konularında daha ileri seviyedeki kavramları öğrenip uygulayabilmek artık sadece zaman meselesi olarak durmaktadır.

# 6. SONUÇ ve ÖNERİLER

Bu tez çalışması kapsamında geliştirdiğimiz DeepResponse-RF yöntemiyle, çoklu omik verinin ve ilaç tanımlayıcı parmak izlerinin bir arada kullanılmasıyla kanser hücre hatlarının ilaç yanıtı tahmin modellemeleri gerçekleştirilmiştir.

Ablasyon analizinde, GDSC verisi kullanılarak her tekil omik veri ve birleştirilmiş omikler veri tipi için özel tahmin modelleri oluşturulmuştur. Analiz sonuçlarında, literatürdeki araştırmalarla örtüşecek şekilde gen ifade verisinin yüksek tahmin performansı getirdiği görülmüştür. Dört farklı omik verinin bir arada kullanan birleştirilmiş omikler verisinin ise gen ifade performansının da üstünde tahmin performansı sunmuştur. Böylelikle, farklı omik veri tiplerinin tahmin performansını iyileştirici etkisi, yöntemimizle gerçekleştirmeyi hedeflediğimiz amaçları destekler nitelikte olmuştur.

Veri içi alan analizinde, GDSC'nin dokulara ayrılan veri yapıları üzerinde üç bölümlendirme stratejisi ve dokuz farklı RF hiperparametre kombinasyonu uygulanarak tahmin performansları ölçülmüştür. Dokulardan elde edilen tahmin performanslarının her metrik için hesaplanan ortalama değerleri incelendiğinde RB tekniğinin diğer bölümlendirme tekniklerine göre daha iyi tahmin performansı getirdiği görülmüştür.

Benzer iki analiz kolu olarak yürütülen GDSC – CCLE ve GDSC – NCI-60 arası çapraz alan analizlerinde, eğitim (GDSC) ve test verisi (CCLE veya NCI-60) arasındaki hücre hattı ve ilaç ortaklıklarını temel alan durumlar için yedi farklı modelleme senaryosu tasarlanmıştır. Her iki analiz kolunda da test verisinin ortak ilaçlara göre düzenlendiği 5 numaralı senaryoya ait tahmin performansı diğer senaryolara göre tüm metriklerde üstün gelmiştir.

DeepResponse-RF'nin ile üretilen ilaç yanıtı tahminlerinin ODTÜ KanSiL Lab tarafından *in vitro* doğrulamaları gerçekleştirilmiştir. Yapılan deneyler, eprinomectin ilaç molekülünün (tahminlerle örtüşecek şekilde) HSK hücre hatları üzerinde uygulanabilecek

yüksek potansiyelli inhibitör özelliği taşıdığını göstermiştir. Alınan bu sonuçlar, yöntemimizin ilaç yeniden konumlandırma aşamaları için uygun bir yöntem olduğunu desteklemiştir.

İleride DeepResponse'de değerlendirilecek veri tiplerinin ayrıntılı analizleri yapılmasıyla önemli kazanımlar elde edileceği düşüncesindeyiz. Özellikle, hücre hattı özellik verisinin değerlendirilmesiyle çeşitli gen kombinasyonlarının önemliliğinin hesaplanması için ek bir analiz yapılabilir. Bu analiz sonucuyla ilaç yanıtını etkileyen biyobelirteçlerin çıkartılması aşamaları oldukça kolaylaşacaktır.

Anti kanser ilaçlarının hedefe yönelik olarak seçilmeleri çevre dokulardaki hasarı azaltabilmek açısından önem taşımaktadır. Bu anlamda, yöntemimizde bulunan kanser hücre hatlarının yanı sıra sağlıklı hücre hatlarının omik veri tipleri ve ilaç yanıtı verileri de gelecekte değerlendirilebilir. Farklı moleküler özelliklere sahip bu hücre hatlarının beraber analiz edilmesiyle kanser hücre hattında yüksek; çevre dokulardaki sağlıklı olanlarda ise düşük etkiye sahip ilaçların seçilmesi mümkün olacaktır.

DeepResponse-RF'nin tasarlanmasındaki uygulamalar bütünü ve tahmin performansları beraber değerlendirildiğinde yöntemin bazı kısıtlayıcı yönleri karşımıza çıkmaktadır. Bunların başlıcaları, GDSC verisinin uzun süren veri ön işleme, modelleme aşamalarındaki zorluklar ve standardizasyon problemleridir. Verideki hücre hattı ve ilaç veri tiplerinin oluşturulup modellerde girdi olarak kullanılabilecek vektörler haline getirilmesi uzun zaman gerektirmektedir. Buradaki zorluk, tüm verinin filtrelenmeden kullanılmasından öte gelmektedir. Bundan dolayı, veri içi alan analizinde GDSC verisinin doku temelli dosyalara bölünüp L1000 gen listesi yardımıyla vektör uzunlukları kısaltılmıştır. Çapraz alan analizinde farklı hücre hattı panellerine ait veri yapılarının bir arada değerlendirilmesi için hücre hattı ve ilaç isimlerinin ortaklıkları belirlenmiş, hücre hattı özellik vektörlerinin uzunlukları eşitlenmiştir. Ancak, panellerin omik veri tiplerini ve deneysel ilaç yanıtı değerlerini elde etmede uyguladıkları yöntemlerin farklı olması

modelleme senaryolarındaki tahmin performanslarını önemli derecede etkilemiştir. Bu standardizasyon probleminden dolayı, GDSC verisi ile eğitilen modellerin CCLE ve NCI-60 test verileri üzerinden oluşturduğu tahminlerde genelleme kabiliyetinin azaldığı görülmektedir.

Yukarıda bahsedilen yetersizlik durumlarının aşılması için teknik açıdan bazı iyileştirmeler yapılabilir. GDSC verisinin doku temelli ayrılmayıp bütünsel olarak değerlendirilmesi için daha güçlü donanım araçları kullanılabilir. Buna ek olarak, modelleme sırasında verinin etkin şekilde kullanılmasına yardımcı olabilecek algoritmalardan (bir Python kütüphanesi olan Dask vb.) yararlanılabileceğini düşünmekteyiz. DeepResponse-RF yönteminin tahmin performansının artırılması amacıyla farklı veri tipleri yeni değişkenler olarak değerlendirme kapsamına alınabilir. Bu yolla, hücrenin moleküler özellikleri arasındaki ilişkiler daha sağlıklı şekilde kurulabilecek ve ilaçların etkileştiği yapılar üzerinden daha açık çıkarımlar yapılabilecektir. Böylelikle, elde edilecek tahminlerin daha güvenilir, klinik öncesi aşamalarda kullanılmaya daha uygun hale geleceğini düşünmekteyiz.

## 7. KAYNAKLAR


1. Weber WW. Pharmacogenetics: From Description to Prediction. Clin Lab Med. 2008;28(4):499-511.

2. Carr DF, Alfirevic A, Pirmohamed M. Pharmacogenomics: Current State-of-the-Art. Genes. 2014;5(2):430-43.

3. Ulukaya E, Karakas D, Dimas K. Tumor Chemosensitivity Assays Are Helpful for Personalized Cytotoxic Treatments in Cancer Patients. Medicina (Mex). 2021;57(6):636.

4. Cook J, Hunter G, Vernon JA. The Future Costs, Risks and Rewards of Drug Development. PharmacoEconomics. 2009;27(5):355-63.

5. Flowers CR, Veenstra D. The Role of Cost-Effectiveness Analysis in the Era of Pharmacogenomics. PharmacoEconomics. 2004;22(8):481-93.

6. Dere WH, Suto TS. The role of pharmacogenetics and pharmacogenomics in improving translational medicine. Clin Cases Miner Bone Metab Off J Ital Soc Osteoporos Miner Metab Skelet Dis. 2009;6(1):13-6.

7. Wang D, Hensman J, Kutkaite G, Toh TS, Galhoz A, GDSC Screening Team, vd. A statistical framework for assessing pharmacological responses and biomarkers using uncertainty estimates. eLife. 04 Aralık 2020;9:e60352.

8. Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A. Machine learning approaches to drug response prediction: challenges and recent progress. Npj Precis Oncol. Aralık 2020;4(1):19.

9. Brooks EA, Galarza S, Gencoglu MF, Cornelison RC, Munson JM, Peyton SR. Applicability of Drug Response Metrics for Cancer Studies using Biomaterials. bioRxiv. 2018;408583.

10. Sebaugh JL. Guidelines for accurate EC50/IC50 estimation. Pharm Stat. Mart 2011;10(2):128-34.

11. Gadagkar SR, Call GB. Computational tools for fitting the Hill equation to dose–response curves. J Pharmacol Toxicol Methods. Ocak 2015;71:68-76.

12. Berrouet C, Dorilas N, Rejniak KA, Tuncer N. Comparison of drug inhibitory effects ($IC_{50}$) in monolayer and spheroid cultures [Internet]. Systems Biology; 2020 May [a.yer 04 Temmuz 2022]. Erişim adresi: http://biorxiv.org/lookup/doi/10.1101/2020.05.05.079285

13. Sharma SV, Haber DA, Settleman J. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. Nat Rev Cancer. 2010;10(4):241-53.

14. Ferreira D, Adega F, Chaves R. The Importance of Cancer Cell Lines as in vitro Models in Cancer Methylome Analysis and Anticancer Drugs Testing. Içinde: Lopez-Camarillo C, editör.

Oncogenomics and Cancer Proteomics - Novel Approaches in Biomarkers Discovery and Therapeutic Targets in Cancer [Internet]. InTech; 2013 [a.yer 03 Kasım 2020]. Erişim adresi: http://www.intechopen.com/books/oncogenomics-and-cancer-proteomics-novel-approaches-in-biomarkers-discovery-and-therapeutic-targets-in-cancer/the-importance-of-cancer-cell-lines-as-in-vitro-models-in-cancer-methylome-analysis-and-anticancer-d

15. Parnham MJ, Kricker JA. Factors Determining Plasticity of Responses to Drugs. Int J Mol Sci. 2022;23(4):2068.

16. Welsh M, Mangravite L, Medina MW, Tantisira K, Zhang W, Huang RS, vd. Pharmacogenomic Discovery Using Cell-Based Models. Pharmacol Rev. 2009;61(4):413-29.

17. Goodspeed A, Heiser LM, Gray JW, Costello JC. Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. Mol Cancer Res. 2016;14(1):3-13.

18. Juan-Blanco T, Duran-Frigola M, Aloy P. Rationalizing Drug Response in Cancer Cell Lines. J Mol Biol. 2018;430(18):3016-27.

19. Reinhold WC, Varma S, Rajapakse VN, Luna A, Sousa FG, Kohn KW, vd. Using drug response data to identify molecular effectors, and molecular “omic” data to identify candidate drugs in cancer. Hum Genet. 2015;134(1):3-11.

20. Koromina M, Pandi MT, Patrinos GP. Rethinking Drug Repositioning and Development with Artificial Intelligence, Machine Learning, and Omics. OMICS J Integr Biol. 01 Kasım 2019;23(11):539-48.

21. Veenstra DL, Higashi MK, Phillips KA. Assessing the cost-effectiveness of pharmacogenomics. AAPS PharmSci. Eylül 2000;2(3):80-90.

22. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, vd. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41(D1):D955-61.

23. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER, vd. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569(7757):503-8.

24. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer. Ekim 2006;6(10):813-23.

25. Altman RB. Predicting Cancer Drug Response: Advancing the DREAM. Cancer Discov. 2015;5(3):237-8.

26. Kim S, Bae S, Piao Y, Jo K. Graph Convolutional Network for Drug Response Prediction Using Gene Expression Data. Mathematics. 2021;9(7):772.

27. Parca L, Pepe G, Pietrosanto M, Galvan G, Galli L, Palmeri A, vd. Modeling cancer drug response through drug-specific informative genes. Sci Rep. 2019;9(1):15222.

28. Piyawajanusorn C, Nguyen LC, Ghislat G, Ballester PJ. A gentle introduction to understanding preclinical data for cancer pharmaco-omic modeling. Brief Bioinform. 06 Ağustos 2021;bbab312.

29. Chen J, Zhang L. A Survey and Systematic Assessment of Computational Methods for Drug Response Prediction. bioRxiv. 2019;697896.

30. Meyer UA, Zanger UM, Schwab M. Omics and Drug Response. Pharmacol Toxicol. 2013;53(1):475-502.

31. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, vd. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. Kasım 2017;171(6):1437-1452.e17.

32. Kang S, Hong S. Prediction of personalized drugs based on genetic variations provided by DNA sequencing technologies. Genes Genomics. 2011;33(6):591-603.

33. Zhou Y, Arribas GH, Turku A, Jürgenson T, Mkrtchian S, Krebs K, vd. Rare genetic variability in human drug target genes modulates drug response and can guide precision medicine. Sci Adv. 2021;7(36):eabi6856.

34. Naulaerts S, Dang CC, Ballester PJ. Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumours. Oncotarget. 2014;5(0):97025-40.

35. Levatić J, Salvadores M, Fuster-Tormo F, Supek F. Mutational signatures are markers of drug sensitivity of cancer cells. bioRxiv. 2021;2021.05.19.444811.

36. Shen L, Kondo Y, Ahmed S, Boumber Y, Konishi K, Guo Y, vd. Drug Sensitivity Prediction by CpG Island Methylation Profile in the NCI-60 Cancer Cell Line Panel. Cancer Res. 2007;67(23):11335-43.

37. Gomez A, Ingelman-Sundberg M. Pharmacoepigenetics: Its Role in Interindividual Differences in Drug Response. Clin Pharmacol Ther. Nisan 2009;85(4):426-30.

38. Miranda SP, Baião FA, Fleck JL, Piccolo SR. Predicting drug sensitivity of cancer cells based on DNA methylation levels. PLOS ONE. 2021;16(9):e0238757.

39. Teodoridis JM, Strathdee G, Plumb JA, Brown R. CpG-island methylation and epigenetic control of resistance to chemotherapy. Biochem Soc Trans. 2004;32(6):916-7.

40. Maier S, Dahlstroem C, Haefliger C, Plum A, Piepenbrock C. Identifying DNA Methylation Biomarkers of Cancer Drug Response. Am J Pharmacogenomics. 2005;5(4):223-32.

41. Vural S, Palmisano A, Reinhold WC, Pommier Y, Teicher BA, Krushkal J. Association of expression of epigenetic molecular factors with DNA methylation and sensitivity to chemotherapeutic agents in cancer cell lines. Clin Epigenetics. 2021;13(1):49.

42. Cascorbi I, Schwab M. Epigenetics in Drug Response. Clin Pharmacol Ther. 2016;99(5):468-70.

43. He Y, Hoskins JM, McLeod HL. Copy number variants in pharmacogenetic genes. Trends Mol Med. 2011;17(5):244-51.

44. Hampton T. Disease, Drug Response Linked to Loss or Gain of Big DNA Chunks in Genome. JAMA. 2007;297(14):1539-40.

45. Gamazon ER, Huang RS, Dolan ME, Cox NJ. Copy number polymorphisms and anticancer pharmacogenomics. Genome Biol. 2011;12(5):R46-R46.

46. Willyard C. Copy number variations' effect on drug response still overlooked. Nat Med. 2015;21(3):206-206.

47. Manica M, Oskooei A, Born J, Subramanian V, Sáez-Rodríguez J, Rodríguez Martínez M. Toward Explainable Anticancer Compound Sensitivity Prediction via Multimodal Attention-Based Convolutional Encoders. Mol Pharm. 02 Aralık 2019;16(12):4797-806.

48. Huang EW, Bhope A, Lim J, Sinha S, Emad A. Tissue-guided LASSO for prediction of clinical drug response using preclinical samples. PLOS Comput Biol. 2020;16(1):e1007607.

49. Zhang L, Chen X, Guan NN, Liu H, Li JQ. A Hybrid Interpolation Weighted Collaborative Filtering Method for Anti-cancer Drug Response Prediction. Front Pharmacol. 2018;9:1017.

50. Niepel M, Hafner M, Pace EA, Chung M, Chai DH, Zhou L, vd. Profiles of Basal and Stimulated Receptor Signaling Networks Predict Drug Response in Breast Cancer Lines. Sci Signal [Internet]. 24 Eylül 2013 [a.yer 08 Temmuz 2022];6(294). Erişim adresi: https://www.science.org/doi/10.1126/scisignal.2004379

51. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, vd. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. PLoS ONE. 2013;8(4):e61318.

52. Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics. 2019;35(14):i501-9.

53. Liu Q, Hu Z, Jiang R, Zhou M. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics. 2020;36(Supplement_2):i911-8.

54. Cortés-Ciriano I, van Westen GJP, Bouvier G, Nilges M, Overington JP, Bender A, vd. Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. Bioinformatics. 08 Eylül 2015;btv529.

55. Rajapakse VN, Luna A, Yamade M, Loman L, Varma S, Sunshine M, vd. CellMinerCDB for Integrative Cross-Database Genomics and Pharmacogenomics Analyses of Cancer Cell Lines. iScience. 2018;10:247-64.

56. Home page - Cancerrxgene - Genomics of Drug Sensitivity in Cancer [İnternet]. [a.yer 30 Haziran 2022]. Erişim adresi: https://www.cancerrxgene.org/

57. Help and Documentation - Cancerrxgene - Genomics of Drug Sensitivity in Cancer [İnternet]. [a.yer 26 Kasım 2020]. Erişim adresi: https://www.cancerrxgene.org/help#t_curve

58. Rolón M, Vega C, Escario JA, Gómez-Barrio A. Development of resazurin microtiter assay for drug sensibility testing of Trypanosoma cruzi epimastigotes. Parasitol Res. Temmuz 2006;99(2):103-7.

59. DepMap Data Downloads [İnternet]. [a.yer 04 Haziran 2021]. Erişim adresi: https://depmap.org/portal/download/

60. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, vd. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 29 Mart 2012;483(7391):603-7.

61. Boyd MR. The NCI Human Tumor Cell Line (60-Cell) Screen. Içinde: Teicher BA, Andrews PA, editörler. Anticancer Drug Development Guide [Internet]. Totowa, NJ: Humana Press; 2004 [a.yer 22 Şubat 2022]. s. 41-61. Erişim adresi: http://link.springer.com/10.1007/978-1-59259-739-0_3

62. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, vd. A gene expression database for the molecular pharmacology of cancer. Nat Genet. 2000;24(3):236-44.

63. NCI-60 Growth Inhibition Data - NCI DTP Data - NCI Wiki [İnternet]. [a.yer 01 Temmuz 2022]. Erişim adresi: https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-60+Growth+Inhibition+Data

64. Shankavaram UT, Varma S, Kane D, Sunshine M, Chary KK, Reinhold WC, vd. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. BMC Genomics. 2009;10(1):277.

65. Rajapakse VN, Luna A, Yamade M, Loman L, Varma S, Sunshine M, vd. Integrative analysis of pharmacogenomics in major cancer cell line databases using CellMinerCDB. bioRxiv. 2018;292904.

66. Luna A, Elloumi F, Varma S, Wang Y, Rajapakse VN, Aladjem MI, vd. CellMiner Cross-Database (CellMinerCDB) version 1.2: Exploration of patient-derived cancer cell line pharmacogenomics. Nucleic Acids Res. 08 Ocak 2021;49(D1):D1083-93.

67. O'Boyle NM. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. J Cheminformatics. Aralık 2012;4(1):22.

68. Ucak UV, Ashyrmamatov I, Lee J. Reconstruction of lossless molecular representations, SMILES and SELFIES, from fingerprints [Internet]. Chemistry; 2022 Haz [a.yer 24 Haziran 2022]. Erişim adresi: https://chemrxiv.org/engage/chemrxiv/article-details/62a1675a804dbe75f63f8ec1

69. Bjerrum EJ. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. ArXiv170307076 Cs [Internet]. 17 Mayıs 2017 [a.yer 05 Mart 2022]; Erişim adresi: http://arxiv.org/abs/1703.07076

70. Rhea help results [İnternet]. [a.yer 13 Temmuz 2022]. Erişim adresi: https://www.rhea-db.org/help/smiles

71. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, vd. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 08 Ocak 2021;49(D1):D1388-95.

72. Wishart DS. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 01 Ocak 2006;34(90001):D668-72.

73. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. J Comput Aided Mol Des. Ağustos 2016;30(8):595-608.

74. Zagidullin B, Wang Z, Guan Y, Pitkänen E, Tang J. Comparative analysis of molecular fingerprints in prediction of drug combination effects. Brief Bioinform. 2021;22(6):bbab291.

75. Rogers D, Hahn M. Extended-Connectivity Fingerprints. J Chem Inf Model. 24 Mayıs 2010;50(5):742-54.

76. Capecchi A, Probst D, Reymond JL. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Cheminformatics. Aralık 2020;12(1):43.

77. Morgan HL. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. J Chem Doc. 01 Mayıs 1965;5(2):107-13.

78. Landrum G, Tosco P, Kelley B, Ric, Sriniker, Gedeck, vd. rdkit/rdkit: 2022_03_4 (Q1 2022) Release [İnternet]. Zenodo; 2022 [a.yer 13 Temmuz 2022]. Erişim adresi: https://zenodo.org/record/591637

79. useR! Machine Learning Tutorial [İnternet]. [a.yer 13 Temmuz 2022]. Erişim adresi: https://koalaverse.github.io/machine-learning-in-R/random-forest.html

80. Shah SH, Angel Y, Houborg R, Ali S, McCabe MF. A Random Forest Machine Learning Approach for the Retrieval of Leaf Chlorophyll Content in Wheat. Remote Sens. 16 Nisan 2019;11(8):920.

81. Breiman L. Random Forest. Mach Learn. 2001;45(1):5-32.

82. Combining Bagging and Random Subspaces to Create Better [İnternet]. [a.yer 14 Temmuz 2022]. Erişim adresi: https://slidetodoc.com/combining-bagging-and-random-subspaces-to-create-better/

83. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. Second edition. New York NY: Springer; 2021. 607 s. (Springer texts in statistics).

84. Rahman R, Matlock K, Ghosh S, Pal R. Heterogeneity Aware Random Forest for Drug Sensitivity Prediction. Sci Rep. 2017;7(1):11347.

85. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, vd. Scikit-learn: Machine Learning in Python. 2012 [a.yer 14 Temmuz 2022]; Erişim adresi: https://arxiv.org/abs/1201.0490

86. Reback J, Jbrockmendel, McKinney W, Van Den Bossche J, Roeschke M, Augspurger T, vd. pandas-dev/pandas: Pandas 1.4.3 [İnternet]. Zenodo; 2022 [a.yer 15 Temmuz 2022]. Erişim adresi: https://zenodo.org/record/3509134

87. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, vd. Array programming with NumPy. Nature. 17 Eylül 2020;585(7825):357-62.

88. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, vd. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 02 Mart 2020;17(3):261-72.

89. Swain M. PubChemPy [İnternet]. 2022 [a.yer 15 Temmuz 2022]. Erişim adresi: https://github.com/mcs07/PubChemPy

90. Kim S, Thiessen PA, Cheng T, Yu B, Bolton EE. An update on PUG-REST: RESTful interface for programmatic access to PubChem. Nucleic Acids Res. 02 Temmuz 2018;46(W1):W563-70.

91. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007;9(3):90-5.

92. Waskom M. seaborn: statistical data visualization. J Open Source Softw. 06 Nisan 2021;6(60):3021.

93. Maaten L van der, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9(86):2579-605.

94. Roberts J, Crall J. alan-turing-institute/distinctipy: v1.2.2 [İnternet]. Zenodo; 2022 [a.yer 17 Temmuz 2022]. Erişim adresi: https://zenodo.org/record/6803948

95. Chen Y, Zhang L. How much can deep learning improve prediction of the responses to drugs in cancer cell lines? Brief Bioinform. 17 Ocak 2022;23(1):bbab378.

96. Zuo Z, Wang P, Chen X, Tian L, Ge H, Qian D. SWnet: a deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures. BMC Bioinformatics. 2021;22(1):434.

97. Wang L, Li X, Zhang L, Gao Q. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. BMC Cancer. 2017;17(1):513.

98. Ma T, Liu Q, Li H, Zhou M, Jiang R, Zhang X. DualGCN: a dual graph convolutional network model to predict cancer drug response. BMC Bioinformatics. Nisan 2022;23(S4):129.

99. Cichonska A, Pahikkala T, Szedmak S, Julkunen H, Airola A, Heinonen M, vd. Learning with multiple pairwise kernels for drug bioactivity prediction. Bioinformatics. 01 Temmuz 2018;34(13):i509-18.

100. Guan NN, Zhao Y, Wang CC, Li JQ, Chen X, Piao X. Anticancer Drug Response Prediction in Cell Lines Using Weighted Graph Regularized Matrix Factorization. Mol Ther - Nucleic Acids. 2019;17:164-74.

101. Xia F, Allen J, Balaprakash P, Brettin T, Garcia-Cardona C, Clyde A, vd. A cross-study analysis of drug response prediction in cancer cell lines. ArXiv210408961 Q-Bio [Internet]. 13 Ağustos 2021 [a.yer 28 Ağustos 2021]; Erişim adresi: http://arxiv.org/abs/2104.08961

102. Hoffman RM. In vitro sensitivity assays in cancer: A review, analysis, and prognosis. J Clin Lab Anal. 1991;5(2):133-43.

103. McDermott U. Cancer cell lines as patient avatars for drug response prediction. Nat Genet. 2018;50(10):1350-1.

104. Pasch CA, Favreau PF, Yueh AE, Babiarz CP, Gillette AA, Sharick JT, vd. Patient-Derived Cancer Organoid Cultures to Predict Sensitivity to Chemotherapy and Radiation. Clin Cancer Res. 01 Eylül 2019;25(17):5376-87.

105. Targeted Therapies - My Cancer Genome [İnternet]. [a.yer 11 Eylül 2022]. Erişim adresi: https://www.mycancergenome.org/content/page/overview-of-targeted-therapies-for-cancer/

106. HGVS recommendations: protein sequence variants [İnternet]. [a.yer 14 Aralık 2020]. Erişim adresi: https://www.hgvs.org/mutnomen/recs-prot.html

107. ExPASy - Cellosaurus [İnternet]. [a.yer 17 Aralık 2020]. Erişim adresi: https://web.expasy.org/cellosaurus/

# 8. EKLER

## 8.1. EK-1

### 8.1.1. Ablasyon Analizi İçin Seçilen Veri Tipleri ve Özellikleri

GDSC'ye ait olan ve alan analizleri için üretilmiş hücre hattı özellik veri tipleri kullanılarak oluşturulan senaryolarda her veri tipi için ve tüm veri tiplerinin bir arada olduğu birleştirilmiş tablo için ayrı ayrı analizler gerçekleştirilmiştir.

**Tablo 8.1.** Ablasyon analizi için kullanılan veri tipleri ve özellikleri.

| Veri Tipi | Gen sayısı | Hücre hattı sayısı | Tablo boyutu |
|---|---|---|---|
| Gen ifade | 897 | 988 | 988, 899 |
| Mutasyon | 963 | 988 | 988, 964 |
| Metilasyon | 955 | 988 | 988, 956 |
| KSD | 932 | 988 | 988, 933 |
| Birleştirilmiş | 3747 | 988 | 988, 3748 |
| İlaç yanıtı | - | 988 | 337761, 6 |

Aşağıda omik veri ve birleştirilmiş omikler veri tipleri için uygulanan analiz aşamaları belirtilmiştir. Bu analizlerin aşamalarında uygulanan yöntemler aynı olup sadece kullanılan değişkenlik gösterdiğinden veri (ilaç yanıtı tablosu hariç) sadece gen ifade verisi üzerinden uygulama anlatılmıştır.

### 8.1.2. Oluşturulan Tüm Veri Tipleri İçin Yapılan Ablasyon Analizi

Pandas ve NumPy kütüphaneleri yüklendikten sonra ilaç yanıtı ve gen ifade veri tipleri yüklenip hücre hattı sayıları listelenip kontrol edilmiştir. Devamında, ilaç yanıtı tablosu bu iki tablonun hücre hattı ismi üzerinden birleştirilebilmesi için hücre hattı ismi sütunu üzerinden alfabetik olarak sıralanmıştır. İlaç ismi ve ilaç parmak izleri listelenip kontrol edilmiştir.

1024 rakam uzunluğundaki İlaç parmak izlerindeki sayıların ayrı ayrı yazılıp listelenmesi gerçekleştirildi. Sonrasında, bu yeni listeler kullanılarak her bir satıra bir parmak izine ait liste eklenerek ve ilk sütunlarda hücre hattı ve ilaç isimleri olacak şekilde yeni bir tablo oluşturuldu. İlaç yanıtı değerleri ise son sütuna eklendi. Bu tablodaki sayı olarak belirtilen sütun isimlerinin her birine gen ifade verisinin sütun uzunluğu eklenip değiştirildi.

Gen ifade verisi ve oluşturulan yeni tablo hücre hattı ismi sütunları baz alınarak birleştirildi. Birleştirilen tablo kopyalanarak modelde eğitim verisi olması verisi için tablodan ilaç, hücre hattı ismi ve pIC50 sütunları çıkarıldı. Etiket verisi olarak ise birleştirilmiş tablonun pIC50 sütunu kullanılmıştır. Oluşturulan iki yeni tablo tek duyarlıklı ve ondalıklı şekilde yeniden kaydedilmiştir. Son olarak, tablolardaki boşluk değeri olup olmadığı kontrol edilmiştir.

Sonuç dosyaları olarak iki dosya belirlenmiştir. İlk dosya için sütun başlarına kullanılan parametreler, MAE, MSE, RMSE, Spearman, Pearson, $R^2$ metriklerinin ortalamasını belirten isimler eklenmiştir. İkinci dosya için ise yine kullanılan parametre, metrik ismi, skor listesi 1 (5 kat çapraz validasyon yapıldığı için ilk beş değer listeye alınmıştır) sütun isimleri eklenmiştir. İki dosya da kaydedilmiştir.

Sklearn kütüphanesine ait metrikler (MAE, MSE, RMSE, $R^2$) ve RandomForestRegressor metotları, cross_validate metotları; Scipy kütüphanesinden spearmanr, pearsonr metotları hazır hale getirilmiştir. Pearson ve Spearman metrikleri cross_validate metodu içinde çalışmadığından her ikisi için de korelasyon sonuçlarını verecek şekilde fonksiyonlar hazırlandı. Sklearn kütüphanesinin make_scorer metodu ile hazırlanan bu fonksiyonlar yardımıyla Pearson ve Spearman metrikleri cross_validate metodu içinde kullanılabilir hale getirilmiş oldu.

5 katlı çapraz doğrulama işlemi yapabilmek için Kfold parametresi olarak n_splits için 5 değeri verildi ve rastgeleliği kontrol etmek için random_state için 2; shuffle için True değerleri verildi.

Rastgele orman regresyon modelinde uygulanacak olan max_depth (karar ağacında uygulanacak maksimum derinlik), n_estimators (orman içinde bulunacak ağaçların sayısı), verbose (koddaki tüm aşamaların açık halde yazdırılması), random_state parametreleri için sırasıyla 81, 100, 4, 2 değerleri verilmiştir.

Cross_validate metoduna yukarıda belirtilen rastgele orman modeli, model eğitiminde kullanılacak veri, etiket verisinin bir liste halindeki durumu, skorlama metrikleri (MAE, MSE, RMSE, Spearman, Pearson, $R^2$), yukarıda parametreleri belirlenen Kfold değişkeni, 6 değeri verilen n_jobs (iş parçacığı), 3 değeri verilen verbose parametreleri eklenmiştir.

Oluşturulan sonuç dosyaları tekrar açılarak elde edilen çıktılar ilgili sütun altına gelecek şekilde yazdırılmıştır. Tüm metrikler için ortalamaların alındığı sonuç dosyasında MAE, MSE, RMSE metrikleri metot sonucu olarak hata değerini ifade etmek için negatif değerler olduğu için "-1" değeriyle çarpılarak dosyaya yazdırıldı. Tüm sonuçlara ait listelerin bulunduğu dosyada ise tüm metriklere ait çıktıların listelenmiş halleri 5 elemanlı bir liste halinde yazdırılmıştır.

**Tablo 8.2.** Ablasyon analizi sonuçlarının bulunduğu dosya isimleri

| **Veri tipi** | **Verilen Dosya ismi** |
|---|---|
| Gen ifade | GDSC_ablation_gexp_ten_fold_cross_validation_v1 |
| | GDSC_ablation_gexp_ten_fold_cross_validation_v1_all_score_lists_v1 |
| Mutasyon | GDSC_ablation_mut_ten_fold_cross_validation_v1 |
| | GDSC_ablation_mut_ten_fold_cross_validation_v1_all_score_lists_v1 |
| Metilasyon | GDSC_ablation_met_ten_fold_cross_validation_v1 |
| | GDSC_ablation_mut_ten_fold_cross_validation_v1_all_score_lists_v1 |
| KSD | GDSC_ablation_cnv_ten_fold_cross_validation_v1 |
| | GDSC_ablation_cnv_ten_fold_cross_validation_v1_all_score_lists_v1 |
| Birleştirilmiş omikler | GDSC_ablation_all_features_ten_fold_cross_validation_v1 |
| | GDSC_ablation_all_features_ten_fold_cross_validation_v1_all_score_lists_v1 |

## 8.2. EK-2

### 8.2.1. GDSC Veri Tiplerinin Düzenlenmesi

Tasarlanan tüm analiz süreçlerinde model girdi verisi olarak kullanılan GDSC veri tipleri için ayrı aşamalardan ilerleyen ön işlem adımları uygulanmıştır.

Hücre hatları için dört ayrı özellik tipi (gen ifade, mutasyon, metilasyon, KSD) kullanılmıştır. İlk aşamada, her özellik tipi için satır başları gen isimleri, sütun başlıkları hücre hattı isimleri ile belirtilen veri tabloları çeşitli kaynak dosyaları kullanılarak oluşturulmuştur. İkinci aşamada, oluşturulan veri tablolarında aynı sayı ve sırada gen ve hücre hatlarının bulunması için her özellik tablosunda yer alan gen ve hücre hatlarının listeleri oluşturulduktan sonra bu listelerin birleşimi yapılmıştır. Özellik tablolarında eksik olan gen isimleri yeni satır eklenerek; eksik hücre hattı isimleri yeni sütunlar eklenerek ve her tablo aynı eksen ve boyutlara sahip hale getirilmiştir. Üçüncü aşamada, her özellik tablosunun istatistikleri değerlendirilip daha az boş değere sahip olan genler seçilip listelenmiştir. Seçilen gen ve hücre hattı listesine göre her özellik tablosunun eksenleri güncellenerek filtrelenmiş özellik tabloları oluşturulmuştur. Dördüncü aşamada ise her hücre hattına ait özellik vektörünü oluşturmak için filtrelenmiş özellik tabloları (gen ifade, mutasyon, metilasyon ve kopya sayısı değişimi sırası ile) alt alta eklenmesi sağlanmış ve birleştirilmiş özellikler vektör tablosu oluşturulmuştur.

**Gen İfade Veri Tipi**

Hücre hatlarına ait gen ifade düzeylerini içeren “Cell_line_RMA_proc_basalExp.txt” dosyasındaki satırlarda 17737 gen ve sütunlarda 1020 hücre hattı bulunmaktadır. Tablo boyutları rapor içinde satır ve sütun sırasıyla (17737, 1020) halinde gösterilecektir. gene_exp_df ismi verilen bu tablodaki GENE_SYMBOLS adlı sütunda değeri olmayan tablo hücreleri belirlenip silinmiştir (toplamda 318 satır). Sonuç tablo boyutları 17419 satır (gen ismi) ve 1020 sütun (hücre hattı ismi) olmuştur.

GENE_SYMBOLS adlı sütun gene_name olarak adlandırılıp GENE_title sütunu silinmiştir. Sütun isimlerindeki sayılar birer COSMIC kimlik numarası olduğundan dolayı bu değerleri kullanarak hücre hatları isimlerine çevirme işlemleri gerçekleştirildi.

Cell_Lines_Details.xlsx (ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/current_release/Cell_Lines_Details.xlsx) dosyasından "COSMIC tissue classification" isimli sayfadaki tablo baz alınarak COSMIC_ID ve Line sütunları kullanılarak "cosmic_id_cell_line_dict_1" sözlüğü oluşturuldu.

methSampleId_2_cosmicIds.xlsx (1029 rows × 20 columns) dosyasındaki cosmic_id sütunundaki değeri olmayan satırlar silindi ve 1023 satır, 20 sütuna sahip tablo oluştu. Ondalıklı değere sahip cosmic_id sütunu tam sayılı hale dönüştürüldü. cosmic_id ve Sample_Name sütunları kullanılarak cosmic_id_cell_line_dict_2 sözlüğü oluşturuldu.

GDSC.Assay2COSMICID.tsv dosyasındaki (1018 rows, 4 columns boyutlarında) COSMIC_ID ve cell_line_name sütunları kullanılarak cosmic_id_cell_line_dict_3 sözlüğü oluşturuldu.

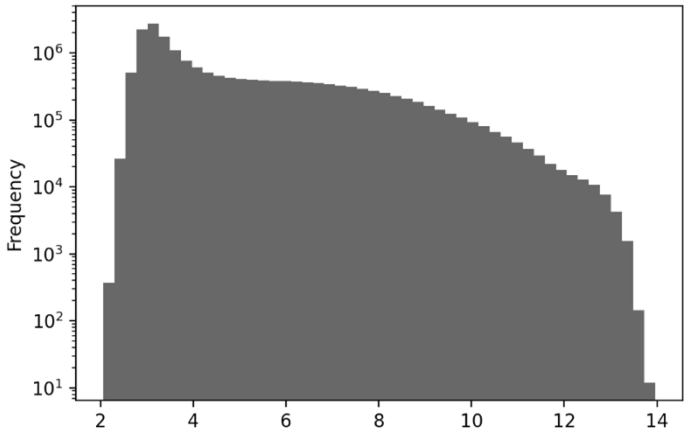
column_name_list adında gene_exp_df tablosunun sütun isimleri listesi oluşturuldu ve "gene_name" elemanı bu listeden çıkarılmıştır. Bu listedeki "Data." ile başlayan sütun isimlerinin bu kısımları silindi. Sütun ismi sonunda ".1" olanların normal olanları (isim sonunda ".1" içermeyenler) halihazırda tablo üzerinde bulunduğundan bu isimler belirlenip (1503362.1, 1330983.1, 909976.1, 905954.1) tablodan silinmiştir ve gene_exp_df sütun sayısı 1015'e inmiştir.

Sütun adlarına karşılık bulmak için oluşturulan üç sözlük ve column_name_list hücre hattı isim listesi kullanıldı. column_name_cell_line_name_dict isimli anahtarları gene_exp_df sütun adları; değerleri açık formdaki hücre hattı isimleri olan bir sözlük oluşturuldu (toplamda 1014 elemanlı). column_name_without_id_list ise eşleşmeyen

sütunları belirtmek için oluşturulmuştu ve tüm sütunlar açık isimle eşleştiği için liste eleman bulundurmamaktadır.

column_name_cell_line_name_dict sözlüğü kullanılarak gene_exp_df sütunları hücre hattı isimleriyle değiştirilmiştir.

Hem gen isimleri hem hücre hatları eksenlerine göre alfabetik sıralama yapılıp sonuçlar GDSC_gene_exp.txt dosyasına yazdırılmıştır.




**Şekil 8.1.** Gen ifade veri tipinin histogram grafiğiyle görselleştirilmesi.

**Mutasyon Veri Tipi**

mutations_20191101.csv dosyası kaynak olarak kullanılmıştır. 1796526 satır, 8 sütun içeren tablodan gene_id, model_id, cancer_driver, rna_mutation, cdna_mutation sütunları silinmiştir. protein_mutation sütunundaki p.0 (hiçbir protein tespit edilemez), p.? (protein analiz edilmedi, bir etki bekleniyor ancak tahmin edilmesi zor) ve "-" değerlerine sahip olan gen bölgeleri kodlanmayan bölge mutasyonu (non-coding mutation) olarak kabul edilmiştir (106). Bu işaretler dışında değerler içeren satırlar ise kodlanan bölge mutasyonu (coding mutation) olarak ele alınmıştır.

Tablonun yeni oluşturulan coding_mutation adlı sütununda p.0, p.?, ve "-" değerleri kodlanmayan bölge mutasyonu oldukları için "0"; başka değer içeren satırlar kodlanan bölge mutasyonu olduklarından "1" değeri verilerek belirtilmiştir.

Yalnızca 'p.0' içeren satırlar ayrı bir noncodingMut_cell_lines1 adlı tabloya aktarıldı ve gene_symbol ve model_name sütunlarına göre aynı değerleri içeren satırlardan sadece birisi tablo üzerinde bırakılarak diğerleri silindi. noncodingMut_cell_lines1_dict1 sözlüğü her bir gen ismi (anahtar) için bir veya daha fazla hücre hattı içeren liste (değer) halinde düzenlenmiştir.

Yalnızca 'p.?' içeren satırlar ayrı bir noncodingMut_cell_lines2 adlı tabloya aktarıldı ve gene_symbol ve model_name sütunlarına göre aynı değerleri içeren satırlardan sadece birisi tablo üzerinde bırakılarak diğerleri silindi. noncodingMut_cell_lines2_dict2 sözlüğü her bir gen ismi (anahtar) için bir veya daha fazla hücre hattı içeren liste (değer) halinde düzenlenmiştir.

Yalnızca “-“ içeren satırlar ayrı bir noncodingMut_cell_lines3 adlı tabloya aktarıldı ve gene_symbol ve model_name sütunlarına göre aynı değerleri içeren satırlardan sadece birisi tablo üzerinde bırakılarak diğerleri silindi. noncodingMut_cell_lines3_dict3 sözlüğü her bir gen ismi (anahtar) için bir veya daha fazla hücre hattı içeren liste (değer) halinde düzenlenmiştir.

Kodlanan bölge mutasyonu içeren satırlar ayrı bir codingMut_cell_lines adlı tabloya aktarıldı ve gene_symbol ve model_name sütunlarına göre aynı değerleri içeren satırlardan sadece birisi tablo üzerinde bırakılarak diğerleri silindi. codingMut_cell_lines_dict sözlüğü her bir gen ismi (anahtar) için bir veya daha fazla hücre hattı içeren liste (değer) halinde düzenlenmiştir. Tablo 1’de her değer için özgün gen ve hücre hattı sayıları verilmiştir.

Her kodlanmayan bölge mutasyonları tablosu için özgün gen ve özgün hücre hattı isimleri ayrı ayrı listeler oluşturuldu. Üç tablodan gelen özgün gen listeleri birleştirilince oluşan listenin eleman sayısı 21285; özgün hücre hattı için oluşan listelerin birleşimi sonrası toplam eleman sayısı ise 1032’dir.

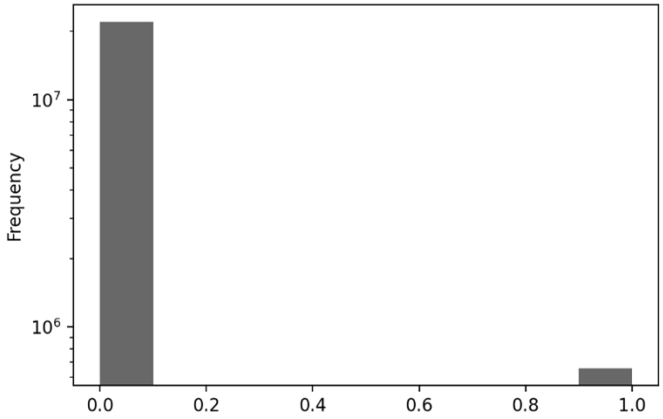
Kodlanan ve kodlanmayan bölge mutasyonlarının özgün gen listeleri birleştirildiğinde 21972 elemanlı mutation_gene_list; özgün hücre hattı listeleri birleştirildiğinde 1032 elemanlı olan kaynak dosyadaki tüm gen ve hücre hatlarını kapsayan mutation_cell_line_list listesi oluşturuldu.

**Tablo 8.3.** Kaynak tablodaki protein_mutation sütunundaki değişkenlere göre özgün gen ve hücre hattı sayıları.

| **Değişkenler** | **Toplam özgün gen sayısı** | **Toplam özgün hücre hattı sayısı** |
|---|---|---|
| p.0 | 22 | 96 |
| p.? | 14458 | 1032 |
| "-" | 20858 | 58 |
| Kodlanan bölge mutasyonları | 19063 | 1032 |

Sonraki adımda, 21972 satır, 1032 sütuna sahip tüm alanlarına "0" değeri atanmış mutation_df tablosu oluşturuldu. Bu tablonun satır başlarına tüm gen isimlerini kapsayan mutation_gene_list; sütun isimlerine ise mutation_cell_line_list listesindeki elemanlar atanmıştır.

mutation_df tablosu üzerindeki kodlanan bölge mutasyonlarını belirtmek için codingMut_cell_lines_dict sözlüğü kullanılarak tablo üzerinde gen ve hücre hattını baz alarak bir arama yapılmış ve eşleşme görülen yerlere "1" değeri atanmıştır. Son hali verilen mutation_df tablosu GDSC_mutation.txt dosyası üzerine yazdırılmıştır.




**Şekil 8.2.** Mutasyon veri tipinin histogram grafiğiyle görselleştirilmesi.

**Metilasyon veri tipi**

Metilasyon veri tipi için ablasyon ve veri içi alan analizinde orijinal veri olan GDSC1000 metilasyon (ikili değerlerden oluşan) veri seti kullanılmıştır.

"PANCAN_methylation_GDSC1000.txt" dosyası kaynak dosya olarak kullanılmıştır. 378 satır, 791 sütundan oluşan tablo methylation_df değişken ismi ile belirtildi. "Unnamed: 0" isimli sütun ismi gene_name olarak değiştirildi. İlk sütun ismi çıkarılarak hücre hatlarına ait cell_line_list adlı liste oluşturuldu. Bu liste içindeki değerler dizi halinden tam sayı haline dönüştürülerek meth_cell_line_list listesine eklendi.

"Cell_list_11112020-gdsc-edited.csv" (https://www.cancerrxgene.org/celllines adresinden alınan dosyanın düzenlenmiş hali) isimli dosyadaki tabo sample_info değişkenine yazılır. Buradaki COSMIC kimlikleri kullanılarak meth_cell_line_list içindeki numaraların hücre hattı karşılıkları elde edilecektir.

Source, COSMIC_ID, Name sütunları cell_line_ids değişkeni içine alındı. COSMIC_ID ve Name sütunlarına göre iki veya daha fazla kopyası olan satırlar silindi. cell_line_dict sözlüğü COSMIC_ID (anahtar) ve Name (değer) sütunları ile oluşturuldu.

meth_data_cell_lines_dict sözlüğü meth_cell_line_list ve cell_line_dict kullanılarak oluşturuldu. meth_data_cell_lines_dict içindeki eleman sayısı 781

bulunmuştur. meth_cell_line_list içindeki eleman sayısı 790 olduğundan dolayı kaynak dosyadaki hücre hattı sayısını yakalamak için eksik 9 (790 – 781) kimliğin açık form hücre hattı ismi Cellosaurus web sitesinden (107) elde edilerek new_ids sözlüğüne eklenir ve meth_data_cell_lines_dict sözlüğü bu yeni sözlükle güncellenir. meth_data_cell_lines_dict kullanılarak methylation_df sütun isimlerine hücre hattı isimleri atandı.

'TableS2H_GDSC1000_meth.xlsx' (https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources//Data/suppData/TableS2H.xlsx) isimli dosyadan genomik koordinatların gen ismi karşılıkları elde edilmeye çalışıldı. Sütun isimlerinden "Genomic Coordinates", "Genomic_Coordinates" ile; "GN" ise "Gene_Name" ile değiştirilmiştir.

Bu iki sütun gcoor_gname isimli değişkene aktarılır ve her iki satıra dayanarak kopya satırlar silindi. gcoor_gname_dict sözlüğü Genomic_Coordinates (anahtar) ve Gene_Name (değer) sütunları ile oluşturuldu. Bu sözlükteki birden çok değer içeren elemanlar “; “ ile ayrıldılar. Tek değere sahip olan anahtarlar gcoor_single_value sözlüğüne aktarıldı. İki veya daha fazla değere sahip olan anahtarlar ise gcoor_multiple_values sözlüğüne eklendi. gcoor_multiple_value_list2 listesi gcoor_multiple_values sözlüğündeki tüm değerlerin bir araya getirilmesiyle oluşturuldu.

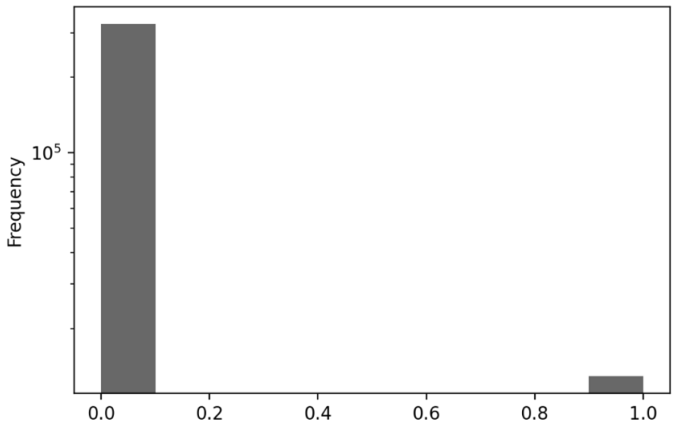
gcoor_single_value sözlüğü tek değerleri içerdiği için methylation_df üzerindeki genomik koordinatları gen isimleri ile değiştirmek için kullanıldı. Değişmeden kalan satırlar multiple_valued_gcoor_rows tablosuna aktarıldı. Bu tablo için indeks yenilendi, gene_name sütununa alfabetik göre sıralama yapıldı ve bu sütun indeks haline getirildi.

cell_line_counts sözlüğü, gcoor_multiple_values sözlüğünün anahtarı ve değer içindeki eleman sayısı ile oluşturuldu ('chr10:102891010-102891794': 2 halinde). gcoor_multple_cell_line_count tablosu için cell_line_counts sözlüğü kullanıldı. İndeks ve “0” sütun isimleri sırasıyla gene_name ve count isimleriyle değiştirildi. Tablo, gene_name sütununa göre alfabetik sıralandı ve gene_name sütunu indeks haline getirildi.

result tablosunda multiple_valued_gcoor_rows ve gcoor_multple_cell_line_count tabloları indeksler baz alınarak birleştirildi (gcoor_multple_cell_line_count count sütunu diğer tablonun en sağına eklendi) ve indeks yenilendi. count sütunundaki değere göre her satırın kopyaları oluşturulup sonuç tablo result_rows_multiplied değişkenine atandı.

gcoor_multiple_value_list2 listesi bir sütuna çevrilip sütuna gene_name2 adı verildi. Tablo, gene_name_column değişkenine atandı. result_rows_multiplied ve gene_name_column sırayla düşey eksende birleştirildi ve tablo result_gene_name_added değişkenine atandı. gene_name2 sütunu ilk sütun yerine atandı ve count, gene_name sütunları silindi. gene_name2 sütun ismi gene_name ile değiştirildi.

methylation_df tablosundan ‘chr’ içeren satırlar (değiştirilmemiş olanlar) silindi. methylation_df ve result_gene_name_added tabloları yatay eksende birleştirildi ve sonuç methylation_table değişkenine atandı.




**Şekil 8.3.** Metilasyon (GDSC1000 verisi) veri tipinin histogram grafiğiyle görselleştirilmesi.

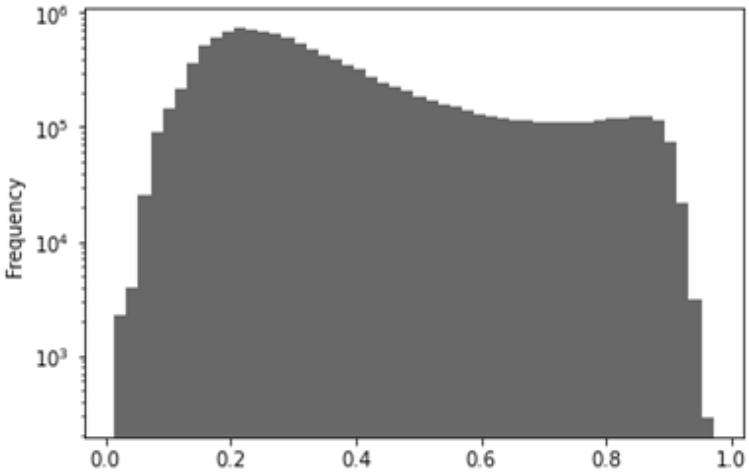
Metilasyon veri tipi için çapraz alan analizinde orijinal veriden (GDSC1000) farklı olarak, beta değerlerine sahip olan CellMinerCDB’den alınan metilasyon veri seti kullanılmıştır.

**Metilasyon Veri Tipi (CellMiner Kaynaklı)**

Pandas, Numpy kütüphaneleri kullanıma hazır hale getirildikten sonra, data_GDSC-MGH-Sanger_met.txt dosyasındaki tablo bir değişkene atanmıştır. Tablodan, genel tablo formatına uygun olmayan üç sütun çıkarılmıştır. Gen ismi sütunu başlığına ait değer yeniden düzenlenmiştir.

Tablodaki hücre hattı ve gen isimleri birer listede tutulmuştur. Gen ismi sütunu indeks haline getirilip tablonun her iki ekseni de alfabetik olarak sıralanmıştır.

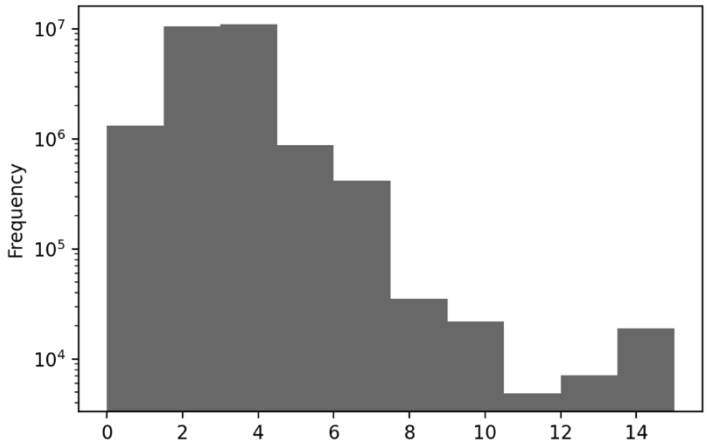
Elde edilen sonuç tablosu bir dosyaya yazdırılıp kaydedilmiştir.




**Şekil 8.4.** Metilasyon (CellMiner verisi) veri tipinin histogram grafiğiyle görselleştirilmesi.

**KSD veri tipi**

cnv_abs_copy_number_picnic_20191101.csv dosyası kaynak olarak kullanılmıştır. İlk işlem olarak, ilk satır silindikten sonra kalan tablodan ilk satır ve model_name sütunu silinip “Unnamed: 1” sütun ismi gene_name ile değiştirilmiştir. Sonuç tablosu GDSC_cnv_picnic.txt dosyasına yazdırılmıştır.




**Şekil 8.5.** KSD veri tipinin histogram grafiğiyle görselleştirilmesi.

**İlaç Yanıtı Veri Tipi**

Veri içi alan analizinde kullanılmak üzere önceden oluşturulan veri (338950 veri noktalı) ile diğer analiz tiplerinde kullanılan veri (337761 veri noktalı) farklı şekilde yapılandırılmıştır. Buna yol açan neden olarak ise farklı isimlere sahip olan aynı ilaçların belirlenerek veriden çıkarılması gösterilebilir.

GDSC'nin iki ayrı deney metodolojisi uygulayarak oluşturduğu GDSC1 ve GDSC2 ilaç yanıtı dosyaları GDSC1_fitted_dose_response_25Feb20.csv ve GDSC2_fitted_dose_response_25Feb20.csv dosyalarındaki tablolar sırasıyla alt alta birleştirilip ikili değerlerden sadece GDSC2 kaynaklı olanların kullanımı önerildiği için tutulmuştur. GDSC1'de yapılan bazı deneyler GDSC2 içinde iyileştirilen ekipman ve prosedürlerle tekrar gerçekleştirildiği için araştırmacılar yalnızca GDSC2'deki sonuçların göz önüne alınması gerektiğini belirmişlerdir (57).

'GDSC1_fitted_dose_response_25Feb20.csv' ve 'GDSC2_fitted_dose_response_25Feb20.csv' dosyaları kaynak dosya olarak kullanılmıştır. Sırasıyla gdsc1_df (310904, 19) ve gdsc2_df (135242, 19) değişkenlerine aktarılan dosyalar sonraki adımda yatay eksende birleştirildi. Sonuç tablosu gdsc1_2_merged_df (446146, 19) değişkenine aktarıldı.

gdsc1_2_merged_df 'DATASET', 'CELL_LINE_NAME', 'DRUG_ID', ‘DRUG_NAME’, 'LN_IC50' sütunları haricindekiler silindi ve 'CELL_LINE_NAME', 'DRUG_ID' sütunlarına göre kopya satırlar bulunup duplicates (52222, 5) adlı değişkene atandı.

Duplicates tablosu 'CELL_LINE_NAME', 'DRUG_ID' sütunlarına göre alfabetik sıralandı. Tablo içinde DATASET sütunundaki değeri GDSC1 olanlar gdsc1_duplicate_df değişkenine aktarıldı. gdsc1_duplicate_df indeks değerleri gdsc1_duplicate_indexes (26111) listesine aktarıldı.

gdsc1_duplicate_indexes tablosundan gdsc1_duplicate_indexes listesindeki indeks numaralarına sahip satırlar silindi ve sonuç result_table (420035, 5) değişkenine yazıldı.

result_table_sorted değişkeni içine result_table DATASET sütunu olmadan aktarıldı. result_table_sorted tablosunda 'DRUG_NAME', 'CELL_LINE_NAME' sütunlarına göre satır bazlı birleştirim yapıldı. Böylece farklı ilaç kimliklerine sahip olan ilaçlar ve buna göre hücre hattı-ilaç çiftleri için farklı olan LN_IC50 değerleri bir araya getirilmiş oldu.

result_table_sorted tablosundaki LN_IC50 sütunu index_ln_ic50_df değişkenine aktarıldı ve bir satır içindeki farklı değerler ayrı sütunlara taşındı. İndeks sütunu tablo içine alınıp ismi index_col gösterildi olarak değiştirildi. Birinci, ikinci, üçüncü LN_IC50 değerleri sırasıyla a,b,c sütunlarında gösterildi.

Sadece 1 LN_IC50 değeri içeren satırlar index_ln_ic50_df’den index_ln_ic50_df_1 değişkenine atandı ve value sütununda seçilen değerler gösterildi.

Sadece 2 LN_IC50 değeri içeren satırlar index_ln_ic50_df’den index_ln_ic50_df_2 değişkenine atandı. Her satırdaki büyük olan LN_IC50 değeri value sütununa yazdırıldı.

Üç farklı LN_IC50 değerine sahip olan satırlar ise index_ln_ic50_df’den index_ln_ic50_df_3 değişkenine atandı. Üç değerin medyan değeri alınıp value sütununda gösterildi.

index_ln_ic50_df_1, index_ln_ic50_df_2, ve index_ln_ic50_df_3 tabloları yatay eksende sırasıyla birleştirilip result adlı değişken üzerine yazdırıldı. result içindeki tablo index_col sütununa göre tablo sıralandı ve bu sütun silindi.

Karşılaştırmak için result_table_sorted içindeki LN_IC50 sütun ismi old_IC50_values ile değiştirildi ve result tablosundaki value sütun ismi LN_IC50 ile değiştirildi. result içindeki LN_IC50 sütunu result_drug_ic50 tablosuna düşey eksende sağ taraftan eklendi.

old_IC50_values sütunu silindi ve IC50 sütunu LN_IC50 sütunundaki değerlerin üstel fonksiyonu alınarak oluşturuldu. Sonrasında, pIC50 sütunu ise IC50 sütunundaki değerlerin 1000000'a bölünüp log10 tabanında hesaplanıp negatif işaretle çarpılarak belirlendi.

Sonuç olarak, result_drug_ic50 tablosu "DRUG_NAME", "CELL_LINE_NAME", "IC50", "LN_IC50", "pIC50" sütun sırasıyla oluşturuldu ve GDSC_drug_response_v3.txt dosyasına yazdırıldı.

Sonuç tablosunda ilaçların SMILES dizi karşılıkları olanları filtrelenmesi için GDSC_drug_smiles_df_v1.txt dosyasındaki tablodan yararlanılmıştır. Tablo, bir değişkene atandıktan sonra SMILES dizisi olmayan ilaçlar tablodan çıkarılmıştır. Kalan ilaç isimleri bir listeye aktarılmıştır. İlaç isimleri, sonuç tablosundan filtrelenmiştir.

Hücre özellik veri tiplerinde bulunan tüm hücre hattı isimleri GDSC_cell_line_name_list.txt dosyasındaki tablodan alınarak bir listeye aktarılmıştır. Listedeki isimler kullanılarak ilaç yanıtı verisindeki hücre hattı isimleri filtrelenmiştir.

Oluşan tablo, ilaç, hücre hattı ismi ve pIC50 sütunlarıyla beraber bir dosyaya yazdırılıp kaydedilmiştir.

Veri içi alan analizi dışındaki analiz tiplerinde kullanılacak olan ilaç yanıtı verisi de yukarıdaki prosedürün bir benzeri izlenerek yapılandırılmıştır. Ancak, ilaç ve hücre hattı isimlerinin diğer platformlarla uygun hale getirilmesi amacıyla düzenlenmesiyle

öncekinden daha küçük bir veri seti elde edilmiştir. Yukarıdaki prosedüre eklenen yeni işlemler aşağıda belirtilmiştir.

Hücre hatları isimlerinde şu düzenlemeler yapılmıştır; "CAPAN-2", "capan2"; "Jurkat", "jurkat"; "SC-1", "sci1". Ardından, düzenlenmiş olan hücre hattı ve ilaç isimleri indekse alınıp tablodaki boşluk değerler silinmiştir. Bu noktadan itibaren uygulanan işlemler dizisi önceki verinin oluşturulmasında uygulananlar ile aynıdır. Sonuç tablosu bir dosyaya kaydedildikten sonra, GDSC_drug_smiles_df_v1.txt dosyası yardımıyla sonuç tablosundan SMILES dizisi olmayan ilaçlar çıkarılmıştır.

Aynı ilacı niteleyen birden çok ismin olması nedeniyle bu ilaçlardan sadece biri tablo üzerinde bırakılmıştır. "sapitinib", "azd8931" arasındaki ve "venetoclax", "venotoclax" isimli ilaçlardan ilk olanları tabloda tutulmuştur, diğerlerine ait satırlar SMILES dizisi olmayan ilaçlarla beraber silinmiştir.

### 8.2.2. Genişletilmiş Veri Tipleri

**Tüm Özellik Dosyalarında Bulunacak Olan Gen Ve Hücre Hattı Listelerinin Oluşturulması**

'GDSC_gene_exp_v2.txt' dosyası gene_exp değişkenine aktarıldı. Bu değişkendeki tablonun sütun isimleri gene_exp_cell_lines (ilk eleman olan gene_name çıkarıldı); satır başları ise gene_exp_genes değişkenine atandı.

'GDSC_cnv_picnic_v1.txt' dosyası cnv_picnic değişkenine aktarıldı. Bu değişkendeki tablonun sütun isimleri cnv_picnic_cell_lines (ilk eleman olan gene_name çıkarıldı); satır başları ise cnv_picnic_genes değişkenine atandı.

'GDSC_mutation_v1.txt' dosyası mutation değişkenine aktarıldı. Bu değişkendeki tablonun sütun isimleri mutation_cell_lines (ilk eleman olan gene_name çıkarıldı); satır başları ise mutation_genes değişkenine atandı.

**Tablo 8.4.** GDSC'den alınan kaynak dosyalarla ilgili bilgiler.

| Veri tipi | Kaynak ve erişim linki | Kaynak veri tablosu boyutu |
|---|---|---|
| Gen ifade | Cell_line_RMA_proc_basalExp.txt https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources//Data/preprocessed/Cell_line_RMA_proc_basalExp.txt.zip | 17737, 1020 |
| Mutasyon | mutations_20191101.csv https://cog.sanger.ac.uk/cmp/download/mutations_20191101.zip | 1796526, 8 |
| Metilasyon | PANCAN_methylation_GDSC1000.txt https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources//Data/BEMs/CellLines/CellLines_METH_BEMs.zip | 378, 791 |
| KSD | cnv_abs_copy_number_picnic_20191101.csv https://cog.sanger.ac.uk/cmp/download/cnv_20191101.zip | 20670, 980 |
| İlaç yanıtı (GDSC1) | GDSC1_fitted_dose_response_25Feb20.csv ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/current_release/GDSC1_fitted_dose_response_25Feb20.xlsx | 310904, 19 |
| İlaç yanıtı (GDSC2) | GDSC2_fitted_dose_response_25Feb20.csv ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/current_release/GDSC2_fitted_dose_response_25Feb20.xlsx | 135242, 19 |

'GDSC_methylation_v1.txt' dosyası methylation değişkenine aktarıldı. Bu değişkendeki tablonun sütun isimleri methylation_cell_lines (ilk eleman olan gene_name çıkarıldı); satır başları ise methylation_genes değişkenine atandı.

'GDSC_drug_response_v2.txt' dosyası drug_response değişkenine aktarıldı. Bu değişkendeki tablonun sütun isimleri drug_response_cell_lines değişkenine atandı.

gene_exp_genes, cnv_picnic_genes, mutation_genes, methylation_genes listelerindeki elemanların birleşim kümesi alındı ve küme alfabetik sıralama ile final_gene_list_2.txt dosyasına yazıldı.

gene_exp_genes, cnv_picnic_genes, mutation_genes, methylation_genes listelerindeki elemanların birleşim kümesi oluşturularak final_gene_list değişkenine atandı ve değişken alfabetik sıralama ile final_gene_list_2.txt dosyasına yazıldı.

gene_exp_cell_lines, cnv_picnic_cell_lines, mutation_cell_lines, methylation_cell_lines, drug_response_cell_lines listelerindeki elemanların birleşim kümesi oluşturularak final_cell_line_list değişkenine atandı ve değişken alfabetik sıralama ile final_cell_line_list_2.txt dosyasına yazıldı.

**Gen İfade Veri Tipi**

'GDSC_gene_exp_v2.txt' dosyası gene_exp değişkenine aktarıldı ve gen ifade tablosundaki gen ve hücre hattı isimleri sırasyla gene_exp_genes ve gene_exp_cell_lines değişkenlerine atandı. Pickle kütüphanesi kullanılarak 'final_gene_list_2.txt' ve 'final_cell_line_list_2.txt' dosyalarındaki isimler sırasıyla final_gene_list, final_cell_line_list adlı listelere aktarılmıştır.

final_gene_list listesinde olup gene_exp_genes'de olmayan isimler new_gene_names listesine yazıldı (toplam 10375 isim).

final_cell_line_list listesinde olup gene_exp_cell_lines'de olmayan isimler new_cell_line_names listesine yazıldı (toplam 112 isim).

gene_exp'in gene_exp_copy adlı kopyası oluşturuldu. Her new_cell_line_names elemanı için gene_exp tablosunda NaN değeri atanmış sütunlar oluşturuldu. gene_exp sütunları liste halinde gene_exp_column_names değişkenine yazıldı.

nan_list adlı boş bir liste oluşturuldu ve her bir gene_exp satırı için NaN içeren tablo hücresi toplamı bu listeye yazıldı. gene_exp üzerinde NaN_count adlı bir sütuna bu listedeki değerler aktarıldı. NaN_count'un her bir satırı toplam hücre hattı değeri olan 1126'ya bölündü ve sonuçlar ratio_NaN adlı sütuna yazıldı.

Tümü NaN değeri atanmış 103375 satır, 1127 sütunluk bir matris blank_dataframe değişkeninde oluşturuldu. Bu matrisin sütun isimleri gene_exp_column_names kullanılarak adlandırıldı. Matristeki gene_name sütunu çıkarıldı ve yeni gene_name sütunu new_gene_names listesindeki elemanlar kullanılarak oluşturuldu.

nan_list2 adlı boş bir liste oluşturuldu ve her bir blank_dataframe satırı için NaN içeren tablo hücresi toplamı bu listeye yazıldı. blank_dataframe üzerinde NaN_count adlı bir sütuna bu listedeki değerler aktarıldı. NaN_count'un her bir satırı toplam hücre hattı değeri olan 1126'ya bölündü ve sonuçlar ratio_NaN adlı sütuna yazıldı.

Aynı sütunlara sahip olan gene_exp ve blank_dataframe tabloları yatay eksende birleştirildi ve sonuç result_table üzerine yazıldı.

result_table üzerindeki değerlerin kontrolünü yapmak için gene_exp_genes ve gene_exp_cell_lines listelerinden rastgele 20'şer eleman seçildi ve gene_exp_copy ve result table tablo hücreleri karşılaştırıldı. Alınan sonuçlara göre hiçbir karşılaştırmada eşleşmeyen durum çıkmadı (False değeri görülmedi).

result_table üzerindeki gene_name, NaN_count, ratio_NaN sütunları nan_result_table değişkenine yazıldı. result_table üzerinden NaN_count, ratio_NaN sütunları silindi.

'GDSC_gene_exp_extended_genes_nan_count_table_v2.txt' adlı dosya içine nan_result_table değişkeni aktarıldı.

result_table üzerindeki gene_name sütunu indeks yapıldı ve her iki eksende alfabetik sıralama yapıldı.

Sütun bazlı NaN içeren tablo hücrelerinin sayıları result_table_cell_line_nan_count isimli değişkene yazıldı. İndeks sütunu tablo içine alınıp ilk ve ikinci sütunlar sırasıyla cell_line ve count isimleriyle adlandırıldı.

'GDSC_gene_exp_nan_count_by_cell_lines_v2.txt' dosyası içine result_table_cell_line_nan_count değişkeni yazdırıldı.

'GDSC_gene_exp_extended_genes_cell_lines_v2.txt' dosyası içine result_table değişkeni yazdırıldı.

**Mutasyon Veri Tipi**

GDSC_mutation_v1.txt dosyası mutation değişkenine aktarıldı ve mutasyon tablosundaki gen ve hücre hattı isimleri sırasyla mutation_genes ve mutation_cell_lines değişkenlerine atandı. Pickle kütüphanesi kullanılarak 'final_gene_list_2.txt' ve 'final_cell_line_list_2.txt' dosyalarındaki isimler sırasıyla final_gene_list, final_cell_line_list adlı listelere aktarılmıştır.

final_gene_list listesinde olup mutation_genes'de olmayan isimler new_gene_names listesine yazıldı (toplam 5822 isim).

final_cell_line_list listesinde olup mutation_cell_lines'de olmayan isimler new_cell_line_names listesine yazıldı (toplam 94 isim).

mutation'nun mutation_copy adlı kopyası oluşturuldu. Her new_cell_line_names elemanı için mutation tablosunda NaN değeri atanmış sütunlar oluşturuldu.

Tümü 0 değeri atanmış 5822 satır, 1127 sütunluk bir matris blank_dataframe değişkeninde oluşturuldu. Bu matrisin sütun isimleri column_names kullanılarak adlandırıldı. Matrisdeki gene_name sütunu çıkarıldı ve yeni gene_name sütunu new_gene_names listesindeki elemanlar kullanılarak oluşturuldu.

mutation'nun mutation_df2 adlı kopyası oluşturuldu. Bu yeni kopyada 0 değerleri NaN ile değiştirildi. zero_list adlı boş bir liste oluşturuldu ve her bir mutation_df2 satırı için NaN içeren tablo hücresi toplamı bu listeye yazıldı. mutation_df2 üzerinde zero_count adlı bir sütuna bu listedeki değerler aktarıldı. NaN_count'un her bir satırı

toplam hücre hattı değeri olan 1126'ya bölündü ve sonuçlar ratio_zeros adlı sütuna yazıldı.

blank_dataframe'nin blank_dataframe_df2 adlı kopyası oluşturuldu. Bu yeni kopyada 0 değerleri NaN ile değiştirildi. zero_list2 adlı boş bir liste oluşturuldu ve her bir blank_dataframe_df2 satırı için NaN içeren tablo hücresi toplamı bu listeye yazıldı. blank_dataframe_df2 üzerinde zero_count adlı bir sütuna bu listedeki değerler aktarıldı. NaN_count'un her bir satırı toplam hücre hattı değeri olan 1126'ya bölündü ve sonuçlar ratio_zeros adlı sütuna yazıldı.

Mutation_df2 ve blank_dataframe_df2'nin 'gene_name', 'zero_count', 'ratio_zeros' adlı sütunları sırasıyla mutation_df3 ve blank_dataframe_df3 değişkenleri üzerine aktarıldı. Bu yeni değişkenler yatay eksende birleştirildi. Sonuç tablosu 'GDSC_mutation_extended_genes_zero_count_table_2.txt' dosyası üzerine yazıldı.

Mutation ve blank_dataframe değişkenleri yatay eksende birleştirildi. Sonuç result_table (27794, 1127) değişkenine aktarıldı.

result_table üzerindeki değerlerin kontrolünü yapmak için mutation_genes ve mutation_cell_lines listelerinden rastgele 20'şer eleman seçildi ve mutation_copy ve result_table tablo hücreleri karşılaştırıldı. Alınan sonuçlara göre hiçbir karşılaştırmada eşleşmeyen durum çıkmadı (False değeri görülmedi).

Result_table'nin indeksi yenilendi ve result_table_copy adlı bir kopyası oluşturuldu. Bu yeni tablodaki tüm 0 değerli alanlar NaN ile değiştirildi. result_table_copy içindeki tüm sütunlara göre NaN içeren satırların kaç tane olduğunun tablosu nan_count değişkenine atandı.

nan_count tablosundaki hücre hattı sütunu tablo içine alındı ve sütun isimleri sırasyıyla cell_line ve count isimleriyle değiştirildi. Tablodaki ilk satır silindi.

nan_count tablosu 'GDSC_mutation_zero_count_by_cell_lines_v2.txt' adlı dosya üzerine yazıldı.

result_table üzerinde gene_name sütunu indeks haline getirildi ve her iki eksende alfabetik sıralama yapıldı.

result_table tablosu 'GDSC_mutation_extended_genes_cell_lines_v2.txt' dosyası üzerine yazıldı.

**Metilasyon Veri Tipi**

GDSC_ methylation_v1.txt dosyası methylation değişkenine aktarıldı ve mutasyon tablosundaki gen ve hücre hattı isimleri sırasyla methylation_genes ve methylation_cell_lines değişkenlerine atandı. Pickle kütüphanesi kullanılarak 'final_gene_list_2.txt' ve 'final_cell_line_list_2.txt' dosyalarındaki isimler sırasıyla final_gene_list, final_cell_line_list adlı listelere aktarılmıştır.

final_gene_list listesinde olup methylation_genes'de olmayan isimler new_gene_names listesine yazıldı (toplam 27367 isim).

final_cell_line_list listesinde olup methylation_cell_lines'de olmayan isimler new_cell_line_names listesine yazıldı (toplam 336 isim).

methylation'nun methylation_copy adlı kopyası oluşturuldu. Her new_cell_line_names elemanı için methylation tablosunda 0 değeri atanmış sütunlar oluşturuldu.

Tümüne 0 değeri atanmış 27367 satır, 1127 sütunluk bir matris blank_dataframe değişkeninde oluşturuldu. Bu matrisin sütun isimleri column_names kullanılarak adlandırıldı. Matristeki gene_name sütunu çıkarıldı ve yeni gene_name sütunu new_gene_names listesindeki elemanlar kullanılarak oluşturuldu.

methylation'nun methylation_df2 adlı kopyası oluşturuldu. Bu yeni kopyada 0 değerleri NaN ile değiştirildi. zero_list adlı boş bir liste oluşturuldu ve her bir methylation_df2 satırı için NaN içeren tablo hücresi toplamı bu listeye yazıldı. methylation_df2 üzerinde zero_count adlı bir sütuna bu listedeki değerler aktarıldı.

NaN_count'un her bir satırı toplam hücre hattı değeri olan 1126'ya bölündü ve sonuçlar ratio_zeros adlı sütuna yazıldı.

blank_dataframe'nin blank_dataframe_df2 adlı kopyası oluşturuldu. Bu yeni kopyada 0 değerleri NaN ile değiştirildi. zero_list2 adlı boş bir liste oluşturuldu ve her bir blank_dataframe_df2 satırı için NaN içeren tablo hücresi toplamı bu listeye yazıldı. blank_dataframe_df2 üzerinde zero_count adlı bir sütuna bu listedeki değerler aktarıldı. NaN_count'un her bir satırı toplam hücre hattı değeri olan 1126'ya bölündü ve sonuçlar ratio_zeros adlı sütuna yazıldı.

Methylation_df2 ve blank_dataframe_df2'nin 'gene_name', 'zero_count', 'ratio_zeros' adlı sütunları sırasıyla methylation_df3 ve blank_dataframe_df3 değişkenleri üzerine aktarıldı. Bu yeni değişkenler yatay eksende birleştirildi. Sonuç tablosu 'GDSC_methylation_extended_genes_zero_count_table_2.txt' dosyası üzerine yazıldı.

Methylation ve blank_dataframe değişkenleri yatay eksende birleştirildi. Sonuç result_table (27794, 1127) değişkenine aktarıldı.

result_table üzerindeki değerlerin kontrolünü yapmak için methylation_genes ve methylation_cell_lines listelerinden rastgele 20'şer eleman seçildi ve methylation_copy ve result_table tablo hücreleri karşılaştırıldı. Alınan sonuçlara göre hiçbir karşılaştırmada eşleşmeyen durum çıkmadı (False değeri görülmedi).

Rresult_table'nin indeksi yenilendi ve result_table_copy adlı bir kopyası oluşturuldu. Bu yeni tablodaki tüm 0 değerli alanlar NaN ile değiştirildi. result_table_copy içindeki tüm sütunlara göre NaN içeren satırların kaç tane olduğunun tablosu nan_count değişkenine atandı.

nan_count tablosundaki hücre hattı sütunu tablo içine alındı ve sütun isimleri sırasyıyla cell_line ve count isimleriyle değiştirildi. Tablodaki ilk satır silindi.

nan_count tablosu 'GDSC_methylation_zero_count_by_cell_lines_v2.txt' adlı dosya üzerine yazıldı.

result_table üzerinde gene_name sütunu indeks haline getirildi ve her iki eksende alfabetik sıralama yapıldı.

result_table tablosu 'GDSC_methylation_extended_genes_cell_lines_v2.txt' dosyası üzerine yazıldı.

**KSD Veri Tipi**

GDSC_cnv_picnic_v1.txt dosyası cnv_picnic değişkenine aktarıldı ve kopya sayısı değişimi tablosundaki gen ve hücre hattı isimleri sırasyla cnv_picnic_genes ve cnv_picnic_cell_lines değişkenlerine atandı. Pickle kütüphanesi kullanılarak 'final_gene_list_2.txt' ve 'final_cell_line_list_2.txt' dosyalarındaki isimler sırasıyla final_gene_list, final_cell_line_list adlı listelere aktarılmıştır.

final_gene_list listesinde olup cnv_picnic_genes'de olmayan isimler new_gene_names listesine yazıldı (toplam 3292 isim).

final_cell_line_list listesinde olup cnv_picnic_cell_lines'de olmayan isimler new_cell_line_names listesine yazıldı (toplam 140 isim).

cnv_picnic'in cnv_picnic_copy adlı kopyası oluşturuldu. Her new_cell_line_names elemanı için cnv_picnic tablosunda 0 değeri atanmış sütunlar oluşturuldu. cnv_picnic sütunları liste halinde cnv_picnic_column_names değişkenine yazıldı.

nan_list adlı boş bir liste oluşturuldu ve her bir cnv_picnic satırı için NaN içeren tablo hücresi toplamı bu listeye yazıldı. cnv_picnic üzerinde NaN_count adlı bir sütuna bu listedeki değerler aktarıldı. NaN_count'un her bir satırı toplam hücre hattı değeri olan 1126'ya bölündü ve sonuçlar ratio_NaN adlı sütuna yazıldı.

Tümü NaN değeri atanmış 3292 satır, 1127 sütunluk bir matris blank_dataframe değişkeninde oluşturuldu. Bu matrisin sütun isimleri cnv_picnic_column_names kullanılarak adlandırıldı. Matristeki gene_name sütunu çıkarıldı ve yeni gene_name sütunu new_gene_names listesindeki elemanlar kullanılarak oluşturuldu.

nan_list2 adlı boş bir liste oluşturuldu ve her bir blank_dataframe satırı için NaN içeren tablo hücresi toplamı bu listeye yazıldı. blank_dataframe üzerinde NaN_count adlı bir sütuna bu listedeki değerler aktarıldı. NaN_count'un her bir satırı toplam hücre hattı değeri olan 1126'ya bölündü ve sonuçlar ratio_NaN adlı sütuna yazıldı.

Aynı sütunlara sahip olan cnv_picnic ve blank_dataframe tabloları yatay eksende birleştirildi ve sonuç result_table (27794, 1129) üzerine yazıldı.

result_table üzerindeki değerlerin kontrolünü yapmak için cnv_picnic_genes ve cnv_picnic_cell_lines listelerinden rastgele 20'şer eleman seçildi ve cnv_picnic_copy ve result_table tablo hücreleri karşılaştırıldı. Alınan sonuçlara göre hiçbir karşılaştırmada eşleşmeyen durum çıkmadı (False değeri görülmedi).

Sütun bazlı NaN içeren tablo hücrelerinin sayıları nan_result_table isimli değişkene yazıldı. İndeks sütunu tablo içine alınıp ilk ve ikinci sütunlar sırasıyla cell_line ve count isimleriyle adlandırıldı ve 0,1127,1128 satırlarındaki elemanlar silindi.

'GDSC_cnv_picnic_nan_count_by_cell_lines_v2.txt' dosyası içine nan_result_table değişkeni yazdırıldı.

result_table üzerindeki gene_name, NaN_count, ratio_NaN sütunları nan_result_table_2 değişkenine yazıldı. result_table üzerinden NaN_count, ratio_NaN sütunları silindi.

'GDSC_cnv_picnic_extended_genes_nan_count_table_v2.txt' adlı dosya içine nan_result_table_2 değişkeni aktarıldı.

result_table üzerindeki gene_name sütunu indeks yapıldı ve her iki eksende alfabetik sıralama yapıldı.

'GDSC_cnv_picnic_extended_genes_cell_lines_v2.txt' dosyası üzerine result_table değişkeni aktarıldı.

**İlaçlar İçin SMILES Dizilerinin Bulunması**

'GDSC_drug_response_v2.txt' adlı dosya result_table değişkenine atandı.

'DrugBank_5.1.6_all_structure_links_SMILES_InChI_Xref.csv' dosyası DrugBank ve SMILES görünümlerini içermektedir. Dosya, df3 değişkenine atandı.

df3 üzerindeki sütun isimleri şöyle değiştirildi; 'DrugBank ID': 'DrugBank_ID', 'PubChem Compound ID': 'PubChem_Compound_ID', 'ChEBI ID':'ChEBI_ID', 'ChEMBL ID': 'ChEMBL_ID'

CHEMBL, CHEBI, PUBCHEM özgün kimliklerinin df3 üzerindeki toplam sayıları sırasıyla 7485, 5047, 8722 şeklindedir.

df3 üzerindeki PubChem_Compound_ID ve ChEBI_ID sütunları float (ondalıklı) halden integer (tam sayı) hale çevrildi.

df3_name_drugbank_id_dict, df3_name_pubchem_id_dict, df3_name_chebi_id_dict, df3_name_chembl_id_dict sözlükleri df3 tablosu baz alınarak anahtar kısımlarında ilaç ismi ve değer kısmında her bir sözlük için sırasıyla DrugBank_ID, PubChem_Compound_ID, ChEBI_ID, ChEMBL_ID sütunlarının değerleri yer almıştır. Her bir sözlük 11405 elemanlıdır.

result_table_drug_name_drug_id_dict sözlüğü result_table tablosunun DRUG_NAME ve DRUG_ID sütunlarıyla oluşturuldu.

DrugBank için df3_name_drugbank_id_dict ve result_table_drug_name_drug_id_dict sözlükleri kullanılarak;

Tam karşılık aramasında drugbank_id_dict_1 içine ilaca karşılık DrugBank kimliği atandı; drugbank_id_matches_dict_1 içine ise her iki tablodaki eşleşen ilaçların isimleri atandı.

Alt metin aramasında drugbank_id_substring_dict_1 içine ilaca karşılık DrugBank kimliği atandı; drugbank_id_substring_matches_dict_1 içine ise her iki tablodaki eşleşen ilaçların isimleri atandı.

Pubchem için df3_name_pubchem_id_dict ve result_table_drug_name_drug_id_dict sözlükleri kullanılarak;

Tam karşılık aramasında pubchem_id_dict_1 içine ilaca karşılık DrugBank kimliği atandı; pubchem_id _matches_dict_1 içine ise her iki tablodaki eşleşen ilaçların isimleri atandı.

Alt metin aramasında pubchem_id _substring_dict_1 içine ilaca karşılık DrugBank kimliği atandı; pubchem_id_substring_matches_dict_1 içine ise her iki tablodaki eşleşen ilaçların isimleri atandı.

result_table üzerindeki 'DRUG_NAME' sütunu baz alınarak drugbank_id_dict_1, df3_name_pubchem_id_dict, df3_name_chembl_id_dict, df3_name_chebi_id_dict sözlükleri sırasıyla 'drugbank_id_1', 'pubchem_id', 'chembl_id', 'chebi_id' aslı yeni sütunlara atandı.

Eşleşen özgün kimlik sayıları yeni sütunlarda drugbank_id_unique_list (139), pubchem_id_unique_list (128), chembl_id_unique_list (138), chebi_id_unique_list (112) şeklindedir.

result_table_drugs_and_drugbank_id_df, result_table içindeki sütunların yeniden sıralanmasıyla oluşturuldu. Bu yeni tabloda sadece GDSC'deki 449 ilaç bırakıldı, diğer satırlar kolaylık açısından dolayı silindi. Pubchem_id ve chebi_id sütunlarındaki boş satırlara NaN değeri atandı.

result_table_drugs_without_drugbank_id_df, result_table_drugs_and_drugbank_id_df içindeki DrugBank kimliği olmayan ilaçların aktarıldığı değişkendir. result_table_drugs_without_drugbank_id_list_1 isimli listesi yeni oluşturulan tablonun özgün DRUGBANK_ID değerlerini içerir ve eleman sayısı 311'dir. Tablonun boyutları ise (311, 7)'dir.

'drugbank vocabulary.csv' dosyasındaki içerik drugbank_vocabulary_df değişkenine aktarıldı.

drugbank_vocabulary_df üzerindeki sütun isimleri şöyle değiştirildi; 'DrugBank ID': 'DrugBank_ID', 'Common name': 'Common_name'

drugbank_vocabulary_df tablosunun drugbank_vocabulary_df_copy adlı kopyası oluşturuldu. Bu kopya tablo üzerinde 'Synonyms_2' adlı sütun Synonyms sütununun içindeki değerlerin "|" işaretlerine göre ayrılıp liste içine koyulması ile oluşturuldu.

drugbank_vocabulary_df_copy_common_name_drugbank_id_dict sözlüğü Common_name ve DrugBank_ID sütunları ile oluşturuldu.

drugbank_vocabulary_df_copy_drugbank_id_synonym_splitted_dict sözlüğü DrugBank_ID ve Synonyms sütunlarıyla oluşturuldu.

drugbank_vocabulary_df_copy_common_name_synonym_splitted_dict sözlüğü Common_name ve Synonyms sütunlarıyla oluşturuldu.

drugbank_vocabulary_df_copy üzerinde 'combined_name_synonyms' isimli sütunu 'Common_name', "|" işareti ve 'Synonyms' sütun değerinin birleştirilmesiyle oluşturuldu.

drugbank_vocabulary_df_copy_combined_name_synonyms_drugbank_id_dict sözlüğü ise combined_name_synonyms ve DrugBank_ID sütunları ile oluşturuldu.

Tam karşılık araması için drugbank_vocabulary_df_copy_common_name_drugbank_id_dict ve

result_table_drug_name_drug_id_dict sözlükleri kullanılarak, drugbank_id_dict_2 içine eşleşen ilaç ve DrugBank kimliği; drugbank_id_matches_dict_2 sözlüğü içine ilaç ve eşleştiği anahtar değeri yazıldı.

Alt metin araması araması için drugbank_vocabulary_df_copy_common_name_drugbank_id_dict ve result_table_drug_name_drug_id_dict sözlükleri kullanılarak, drugbank_id_substring_dict_2 içine eşleşen ilaç ve DrugBank kimliği; drugbank_id_substring_matches_dict_2 sözlüğü içine ilaç ve eşleştiği anahtar değeri yazıldı.

Tam karşılık araması için drugbank_vocabulary_df_copy_combined_name_synonyms_drugbank_id_dict ve result_table_drug_name_drug_id_dict sözlükleri kullanılarak, drugbank_id_dict_3 içine eşleşen ilaç ve DrugBank kimliği; drugbank_id_matches_dict_3 sözlüğü içine ilaç ve eşleştiği anahtar değeri yazıldı.

Alt metin araması araması için drugbank_vocabulary_df_copy_combined_name_synonyms_drugbank_id_dict ve result_table_drug_name_drug_id_dict sözlükleri kullanılarak, drugbank_id_substring_dict_3 içine eşleşen ilaç ve DrugBank kimliği; drugbank_id_substring_matches_dict_3 sözlüğü içine ilaç ve eşleştiği anahtar değeri yazıldı.

'Drug_listFri Nov 27 09_00_17 2020_edited.csv' isimli dosya gdsc_synonym_info_df değişkenine aktarıldı.

drug_name_pubchem_df değişkeni içine gdsc_synonym_info_df tablosunun NAME ve PubCHEM sütunları yazıldı. Yeni tablo içindeki PubCHEM sütunu numerik hale çevrildi ve tablodaki tüm NaN değerleri 0 ile değiştirildi. PubCHEM sütunu tam sayı olarak yeniden oluşturuldu.

drug_name_pubchem_df üzerindeki 0 değerleri NaN ile değiştirildi. PubCHEM sütunu baz alınarak tüm NaN içeren satırlar silindi. Name ve PubCHEM sütunlarına göre kopyalar içeren satırlardan biri bırakılıp diğerleri silindi.

drug_name_pubchem_df üzerinde PubCHEM sütunu tam sayı haline dönüştürüldü. Name sütunundaki ilaçlar baz alınıp PubCHEM kimlikleri bunlara göre satır bazında birleştirme yapıldı.

gdsc_synonym_info_name_pubchem_combined_dict sözlüğü drug_name_pubchem_df tablosunun Name ve PubCHEM sütunları kullanılarak oluşturuldu.

gdsc_synonym_info_df tablosunun Synonyms sütunu baz alınarak boş değer içeren satırlar silindi. Name sütunu baz alınarak farklı kimlikleri bir araya getirmek için satırlar birleştirildi.

gdsc_synonym_info_list listesi gdsc_synonym_info_df içindeki Name sütunu kullanılarak oluşturuldu. gdsc_synonym_info_name_synonym_dict sözlüğü için Name ve Synonyms sütunları kullanıldı.

gdsc_synonym_info_name_synonym_dict sözlüğü değerlerinde birden çok eleman içerenler virgül ile ayrılıp yeniden oluşturuldu.

gdsc_synonym_info_name_synonym_dict sözlüğündeki 'GSK2578215A', 'Olaparib' 'FTY-720', 'Luminespib', 'Tozasertib', 'ML323', 'Daporinad' elemanlarının değerleri tekrar düzenlenip değer olarak atandı.

gdsc_synonym_info_name_pubchem_combined_dict içindeki ikili değer içeren 'BMS-345541' anahtarının bir değeri silindi. Aynı sözlük için her değerin ilk elemanı yeni değer olarak atandı.

gdsc_synonym_info_df tablosu için synonym_2 sütunu Name sütununa göre gdsc_synonym_info_name_synonym_dict sözlüğü kullanılarak oluşturuldu. Benzer

şekilde, pubchem_id_2 sütunu ise gdsc_synonym_info_name_pubchem_combined_dict sözlüğü ile oluşturuldu. gdsc_synonym_info_name_pubchem_id_dict sözlüğü gdsc_synonym_info_df tablosundaki Name ve pubchem_id_2 sütunlarıyla hazırlandı.

gdsc_synonym_info_name_drug_id_dict sözlüğü gdsc_synonym_info_df tablosundaki Name ve drug_id sütunları kullanılarak hazırlandı.

Tam karşılık araması ile gdsc_synonym_info_name_pubchem_combined_dict ve result_table_drug_name_drug_id_dict sözlükleri kullanılarak pubchem_id_dict_2 ve pubchem_id_matches_dict_2 sözlükleri hazırlandı.

Alt metin araması ile gdsc_synonym_info_name_pubchem_combined_dict ve result_table_drug_name_drug_id_dict sözlükleri kullanılarak pubchem_id_substring_dict_2 ve pubchem_id_substring_matches_dict_2 sözlükleri hazırlandı.

result_table_drugs_and_drugbank_id_df tablosunda synonyms sütunu DRUG_NAME sütunu baz alınarak gdsc_synonym_info_name_synonym_dict sözlüğü ile hazırlandı. Diğer yandan, pubchem_id_2 sütunu ise aynı şekilde gdsc_synonym_info_name_pubchem_id_dict sözlüğü ile oluşturuldu.

result_table_drugs_with_drugbank_id_df tablosu result_table_drugs_and_drugbank_id_df tablosundaki drugbank_id_1 sütunundaki dolu sütunlara göre oluşturuldu. Boş olan hücrelerin satırları result_table_drugs_without_drugbank_id_df tablosuna yazıldı.

result_table_drugs_without_drugbank_id_dict sözlüğü result_table_drugs_without_drugbank_id_df tabousndaki DRUG_NAME ve drugbank_id_1 sütunlarına göre hazırlandı.

Tam karşılık aramasında GDSC ilaç isimleri ve drugbank_vocabulary_df'deki DrugBank kimlikleri eşleştirildi. result_table_drugs_without_drugbank_id_dict ve drugbank_vocabulary_df_copy_common_name_drugbank_id_dict sözlükleri

kullanılarak drugbank_id_dict_4 ve drugbank_matches_dict_4 sözlükleri sırasıyla ilaç-drugbank kimliği ve ilaç-eşleşen ilaç içeriği ile hazırlandı.

Alt metin aramasında ise yine aynı sözlüklere göre drugbank_id_substring_dict_4 ve drugbank_id_substring_matches_dict_4 sözlükleri sırasıyla ilaç-drugbank kimliği ve ilaç-eşleşen ilaç içeriği ile hazırlandı.

'drug_annotation_table-2020-11-27T10 40 48-05 00.csv', 'drugs.tsv' ve 'chemicals.tsv' adlı dosyalar sırasıyla drug_table_1, drug_table_2, drug_table_3 değişkenlerine aktarıldı.

Ddrug_table_1 tablosu için sütun isimleri şu şekilde değiştirildi; 'unique.drugid':'unique_drugid', 'GDSC1000.drugid': 'GDSC1000_drug_name', 'cid':'pubchem_id'.

drug_table_1_unique_name_gdsc_name_dict sözlüğü tablodaki unique_drugid ve GDSC1000_drug_name sütunlarıyla oluşturuldu. drug_table_1_gdsc_name_smiles_dict sözlüğü ise GDSC1000_drug_name ve smiles sütunlarıyla oluşturuldu.

Drug_table_1 içindeki 'unique_drugid','pubchem_id' sütunları pubchem_df değişkenine yazıldı. Yeni tabloda pubchem_id sütununa göre boşluk içeren satırlar silindi. Pubchem_id tam sayı olarak yazılarak değiştirildi. 'unique_drugid', 'pubchem_id' sütunlarına göre kopya satırlar silindi. drug_table_1_unique_name_pubchem_id_dict sözlüğü unique_drugid ve pubchem_id sütunlarına göre oluşturuldu. 'GDSC1000_drug_name' sütunu 'unique_drugid' sütunu üzerinden drug_table_1_unique_name_gdsc_name_dict sözlüğü ile hazırlandı. 'GDSC1000_drug_name' sütununa göre boş değer içeren satırlar silindi.

drug_table_1_gdsc_name_pubchem_id_dict sözlüğü GDSC1000_drug_name ve pubchem_id sütunları ile oluşturuldu.

drug_table_2 sütun isimleri şu şekilde değiştirildi; 'Name':'name', 'Generic Names':'generic_names','Cross-references': 'Cross_references', 'PubChem Compound Identifiers': 'PubChem_Compound_Identifiers'.

drug_table_2_name_pubchem_id_dict sözlüğü name ve PubChem_Compound_Identifiers sütunları ile oluşturuldu. drug_table_2_name_smiles_dict sözlüğü ise name ve SMILES sütunları ile hazırlandı.

Re kütüphanesi kullanım için içe aktarıldı. Cross_ref_df tablosu drug_table_2 üzerindeki 'name' sütunu ile hazırlandı. drug_table_2 üzerindeki 'Cross_references' sütunundaki değerler virgüller baz alınarak ayrıldı Cross_ref_df üzerine aktarıldı ve bu sütuna göre boş olan satırlar silindi. Cross_ref_df üzerinde 'pubchem_id' ve 'drugbank_id' sütunları NaN içeriği ile oluşturuldu. Cross_references sütunu içindeki kimlikler pubchem_id ve drugbank_id sütunlarına aktarıldı. Aktarma yapılırken değerler ‘:’ ile ayrıldı ve “”” işareti değer içinden silindi. Cross_ref_df_name_drugbank_id_dict sözlüğü name ve drugbank_id sütunları kullanılarak oluşturuldu. drug_table_2 üzerinde drugbank_id sütunu name sütunu baz alınarak Cross_ref_df_name_drugbank_id_dict sözlüğü ile hazırlandı. drug_table_2_name_drugbank_id_dict sözlüğü drug_table_2 içindeki name ve drugbank_id sütunu ile hazırlandı. Cross_ref_df içindeki drugbank_id sütununa göre boş satırlar silindi.

drug_table_3 sütun isimleri şu şekilde değiştirildi; 'Name': 'name', 'Generic Names': 'generic_names', 'Cross-references': 'Cross_references', 'PubChem Compound Identifiers': 'PubChem_Compound_Identifiers'. drug_table_3_name_pubchem_id_dict sözlüğü name ve PubChem_Compound_Identifiers sütunları ile birlikte oluşturuldu. drug_table_3_name_smiles_dict sözlüğü ise name ve SMILES sütunları ile hazırlandı.

Cross_ref_df_2 tablosu drug_table_3 içindeki name sütunu ile oluşturuldu. drug_table_3 üzerindeki 'Cross_references' sütunundaki değerler virgüller baz alınarak ayrıldı Cross_ref_df_2 üzerine aktarıldı ve bu sütuna göre boş olan satırlar silindi. Cross_ref_df_2 üzerinde 'pubchem_id' ve 'drugbank_id' sütunları NaN içeriği ile

oluşturuldu. Cross_references sütunu içindeki kimlikler pubchem_id ve drugbank_id sütunlarına aktarıldı. Aktarma yapılırken değerler ‘:’ ile ayrıldı ve “”” işareti değer içinden silindi. Cross_ref_df_2_name_drugbank_id_dict sözlüğü name ve drugbank_id sütunları kullanılarak oluşturuldu. drug_table_3 üzerinde drugbank_id sütunu name sütunu baz alınarak Cross_ref_df_2_name_drugbank_id_dict sözlüğü ile hazırlandı. drug_table_3_name_drugbank_id_dict sözlüğü drug_table_3 içindeki name ve drugbank_id sütunu ile hazırlandı. Cross_ref_df_2 içindeki drugbank_id sütununa göre boş satırlar silindi.

drug_table_1 için tam eşleşme araması; PubChem kimliği için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_1_gdsc_name_pubchem_id_dict sözlükleri kullanılarak common_pubchem_id_dict_1 ve common_pubchem_id_matches_dict_1 sözlükleri sırasıyla ilaç-PubChem kimliği ve ilaç-eşleşen ilaç şeklinde hazırlandı.

SMILES için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_1_gdsc_name_smiles_dict sözlükleri kullanılarak common_smiles_dict_1 ve common_smiles_matches_dict_1 sözlükleri sırasıyla ilaç-SMILES ve ilaç-eşleşen ilaç şeklinde hazırlandı.

drug_table_1 için alt metin araması ile; PubChem kimliği için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_1_gdsc_name_pubchem_id_dict sözlükleri kullanılarak common_pubchem_id_substring_dict_1 ve common_pubchem_id_substring_matches_dict_1 sözlükleri sırasıyla ilaç-PubChem kimliği ve ilaç-eşleşen ilaç şeklinde hazırlandı.

SMILES için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_1_gdsc_name_smiles_dict sözlükleri kullanılarak common_smiles_substring_dict_1 ve common_smiles_substring_matches_dict_1 sözlükleri sırasıyla ilaç-SMILES gösterimi ve ilaç-eşleşen ilaç şeklinde hazırlandı.

drug_table_2 için tam karşılık araması ile; DrugBank kimliği için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_2_name_drugbank_id_dict sözlükleri kullanılarak common_drugbank_id_dict_2 ve common_drugbank_id_matches_dict_2 sözlükleri sırasıyla ilaç-drugbank kimliği ve ilaç-eşleşen ilaç şeklinde hazırlandı.

Pubchem kimliği için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_2_name_pubchem_id_dict sözlükleri kullanılarak common_pubchem_id_dict_2 ve common_pubchem_id_matches_dict_2 sözlükleri sırasıyla ilaç-PubChem kimliği ve ilaç-eşleşen ilaç şeklinde hazırlandı.

SMILES gösterimi için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_2_name_smiles_dict sözlükleri kullanılarak common_smiles_dict_2 ve common_smiles_matches_dict_2 sözlükleri sırasıyla ilaç-SMILES gösterimi ve ilaç-eşleşen ilaç şeklinde hazırlandı.

drug_table_2 için alt metin araması ile; DrugBank kimliği için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_2_name_drugbank_id_dict sözlükleri kullanılarak common_drugbank_id_substring_dict_2 ve common_drugbank_id_substring_matches_dict_2 sözlükleri sırasıyla ilaç-drugbank kimliği ve ilaç-eşleşen ilaç şeklinde hazırlandı.

Pubchem kimliği için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_2_name_pubchem_id_dict dict sözlükleri kullanılarak common_pubchem_id_substring_dict_2 ve common_pubchem_id_substring_matches_dict_2 sözlükleri sırasıyla ilaç-PubChem kimliği ve ilaç-eşleşen ilaç şeklinde hazırlandı.

SMILES gösterimi için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_2_name_smiles_dict sözlükleri kullanılarak common_smiles_substring_dict_2 ve common_smiles_substring_matches_dict_2 sözlükleri sırasıyla ilaç-SMILES gösterimi ve ilaç-eşleşen ilaç şeklinde hazırlandı.

drug_table_3 için tam karşılık araması ile; DrugBank kimliği için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_3_name_drugbank_id_dict sözlükleri kullanılarak common_drugbank_id_dict_3 ve common_drugbank_id_matches_dict_3 sözlükleri sırasıyla ilaç-drugbank kimliği ve ilaç-eşleşen ilaç şeklinde hazırlandı.

Pubchem kimliği için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_3_name_pubchem_id_dict sözlükleri kullanılarak common_pubchem_id_dict_3 ve common_pubchem_id_matches_dict_3 sözlükleri sırasıyla ilaç-PubChem kimliği ve ilaç-eşleşen ilaç şeklinde hazırlandı.

SMILES gösterimi için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_3_name_smiles_dict sözlükleri kullanılarak common_smiles_dict_3 ve common_smiles_matches_dict_3 sözlükleri sırasıyla ilaç- SMILES gösterimi ve ilaç-eşleşen ilaç şeklinde hazırlandı.

drug_table_3 için alt metin araması ile; DrugBank kimliği için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_3_name_drugbank_id_dict sözlükleri kullanılarak common_drugbank_id_substring_dict_3 ve common_drugbank_id_substring_matches_dict_3 sözlükleri sırasıyla ilaç-drugbank kimliği ve ilaç-eşleşen ilaç şeklinde hazırlandı.

Pubchem kimliği için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_3_name_pubchem_id_dict sözlükleri kullanılarak common_pubchem_id_substring_dict_3 ve common_pubchem_id_substring_matches_dict_3 sözlükleri sırasıyla ilaç-PubChem kimliği ve ilaç-eşleşen ilaç şeklinde hazırlandı.

SMILES gösterimi için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve drug_table_3_name_smiles_dict sözlükleri kullanılarak common_smiles_substring_dict_3 ve common_smiles_substring_matches_dict_3 sözlükleri sırasıyla ilaç- SMILES gösterimi ve ilaç-eşleşen ilaç şeklinde hazırlandı.

df3 tablosunda 'DrugBank ID':'DrugBank_ID’ şeklinde sütun ismi değişikliği yapıldı. df3_name_smiles_dict sözlüğü df3 üzerindeki Name ve SMILES sütunları ile oluşturuldu. df3_drugbank_id_smiles_dict sözlüğü ise DrugBank_ID ve SMILES sütunları ile hazırlandı.

df3 tablosu için tam karşılık araması ile; SMILES gösterimi için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve df3_name_smiles_dict sözlükleri kullanılarak common_smiles_dict_4 ve common_smiles_matches_dict_4 sözlükleri sırasıyla ilaç-SMILES gösterimi ve ilaç-eşleşen ilaç şeklinde hazırlandı.

df3 tablosu için alt metin araması ile; SMILES gösterimi için result_table_drugs_and_drugbank_id_df_name_synonyms_1_dict ve df3_name_smiles_dict sözlükleri kullanılarak common_smiles_substring_dict_4 ve common_smiles_substring_matches_dict_4 sözlükleri sırasıyla ilaç-SMILES gösterimi ve ilaç-eşleşen ilaç şeklinde hazırlandı.

DrugBank kimliği ile alakalı olan common_drugbank_id_dict_2 ve common_drugbank_id_dict_3 sözlüklerindeki boş değere sahip anahtarların silinmesi için sırasıyla common_drugbank_id_dict_2_nan_removed ve common_drugbank_id_dict_3_nan_removed karşılıkları oluşturuldu.

drugbank_id_dict_1 üzerinde tüm DrugBank sözlüklerini bir araya getirmek için drugbank_id_dict_3, drugbank_id_dict_4, common_drugbank_id_dict_2_nan_removed ve common_drugbank_id_dict_3_nan_removed sözlüğe eklenmiştir.

Pubchem kimliği ile alakalı olan pubchem_id_dict_1, common_pubchem_id_dict_2 ve common_pubchem_id_dict_3 sözlükleri için aynı süreç uygulanarak boş değere sahip anahtarlar silindi. Sözlük öncelikle tablo formatına getirildi. İkinci sütundaki boşluk değerleri silindi ve sütun ismi pubchem_id olarak değiştirildi. Sözlükler için sırasıyla pubchem_id_dict_1_nan_removed, common_pubchem_id_dict_2_nan_removed ve common_pubchem_id_dict_3_nan_removed sözlükleri indeks ve pubchem_id sütunları ile oluşturuldu.

pubchem_id_dict_1_nan_removed üzerinde tüm DrugBank sözlüklerini bir araya getirmek için pubchem_id_dict_2, common_pubchem_id_dict_1, common_pubchem_id_dict_2_nan_removed ve common_pubchem_id_dict_3_nan_removed sözlüğe eklenmiştir.

SMILES gösterimi ile alakalı olan common_smiles_dict_1, common_smiles_dict_2, common_smiles_dict_3 ve common_smiles_dict_4 sözlükleri için aynı süreç uygulanarak boş değere sahip anahtarlar silindi. Sözlük öncelikle tablo formatına getirildi. İkinci sütundaki boşluk değerleri silindi ve sütun ismi SMİLES olarak değiştirildi. Sözlükler için sırasıyla common_smiles_dict_1_nan_removed, common_smiles_dict_2_nan_removed, common_smiles_dict_3_nan_removed ve common_smiles_dict_4_nan_removed sözlükleri indeks ve SMİLES sütunları ile oluşturuldu.

common_smiles_dict_1_nan_removed üzerinde tüm DrugBank sözlüklerini bir araya getirmek için common_smiles_dict_2_nan_removed, common_smiles_dict_3_nan_removed ve common_smiles_dict_4_nan_removed sözlüğe eklenmiştir.

mapped_drugbank_pubchem_smiles_df tablosu result_table_drugs_and_drugbank_id_df tablosundaki 'DRUG_ID', 'DRUG_NAME' 'chembl_id', 'chebi_id' sütunları kullanılarak hazırlandı. 'drugbank_id, 'pubchem_id', 'SMILES' sütunları ise DRUG_NAME sütunu baz alınarak sırasıyla drugbank_id_dict_1, pubchem_id_dict_1_nan_removed ve common_smiles_dict_1_nan_removed sözlükleri kullanılarak oluşturuldu.

**Tablo 8.4.** Alt metin arama eşleşme sözlükleri ve bu sözlüklerin DrugBank kimlik sözlükleri.

| **Alt metin arama eşleşme sözlükleri** | **İlişkili DrugBank kimlik sözlükleri** |
|---|---|
| drugbank_id_substring_matches_dict_1 | df3_name_drugbank_id_dict |
| drugbank_id_substring_matches_dict_3 | drugbank_vocabulary_df_copy_combined_name_synonyms_drugbank_id_dict |
| drugbank_id_substring_matches_dict_4 | drugbank_vocabulary_df_copy_common_name_drugbank_id_dict |
| common_drugbank_id_substring_matches_dict_2 | drug_table_2_name_drugbank_id_dict |
| common_drugbank_id_substring_matches_dict_3 | drug_table_3_name_drugbank_id_dict |

merged_df_drugs_without_drugbank_id_list_from_df tablosu mapped_drugbank_pubchem_smiles_df tablosu içindeki 'drugbank_id' sütununda boş olan satırlar ile oluşturuldu. merged_df_drugs_without_drugbank_id_list listesi ise bu yeni tablodaki DRUG_NAME isimli sütundaki özgün ilaç isimleri ile hazırlandı.

DrugBank kimliği ile alakalı alt metin arama sözlüklerinin düzenlenmesi aşağıdaki gibidir. drugbank_id_substring_matches_dict_1_edited sözlüğü,

drugbank_id_substring_matches_dict_1 sözlüğü üzerindeki anlamsız eşleşmeler çıkarılarak oluşturuldu. drugbank_id_substring_matches_dict_1_edited ve merged_df_drugs_without_drugbank_id_list değişkenleri kullanılarak df3_name_drugbank_id_dict üzerinde tam karşılık araması yapılıp ilaç-drugbank kimliği ve ilaç-eşleşen ilaç ikilileri sırasıyla matched_drugs_with_drugbank_id_substring_dict_1 ve matched_drugs_with_drugbank_id_substring_matches_dict_1 sözlüklerine aktarıldı.

matched_drugs_with_drugbank_id_substring_dict_1 içindeki boş değerli anahtarlar silindi ve matched_drugs_with_drugbank_id_substring_dict_1_nan_removed sözlüğü oluşturuldu.

drugbank_id_substring_matches_dict_3_edited sözlüğü, drugbank_id_substring_matches_dict_3 sözlüğü üzerindeki anlamsız eşleşmeler çıkarılarak oluşturuldu. drugbank_id_substring_matches_dict_3_edited ve merged_df_drugs_without_drugbank_id_list değişkenleri kullanılarak drugbank_vocabulary_df_copy_combined_name_synonyms_drugbank_id_dict üzerinde tam karşılık araması yapılıp ilaç-drugbank kimliği ve ilaç-eşleşen ilaç ikilileri sırasıyla temp ve matched_drugs_with_drugbank_id_substring_matches_dict_2 sözlüklerine aktarıldı.

drugbank_vocabulary_df_copy_combined_name_synonyms_drugbank_id_dict sözlüğündeki değerler “|” işareti ile ayrıldıktan sonra oluşan liste içinde temp sözlüğü kullanılarak drugbank_vocabulary_df_copy_common_name_drugbank_id_dict

üzerinde tam karşılık araması yapıldı. Tek değer içerenlerde ise birebir olarak tam karşılık araması yapıldı. Her iki döngüde de ilaç-drugbank kimliği eşleşmeleri matched_drugs_with_drugbank_id_substring_dict_2 sözlüğüne aktarılmıştır.

common_drugbank_id_substring_matches_dict_2_edited sözlüğü, common_drugbank_id_substring_matches_dict_2 sözlüğü üzerindeki anlamsız

eşleşmeler çıkarılarak oluşturuldu. common_drugbank_id_substring_matches_dict_2_edited

ve merged_df_drugs_without_drugbank_id_list değişkenleri kullanılarak drug_table_2_name_drugbank_id_dict üzerinde tam karşılık araması yapılıp ilaç-drugbank kimliği ve ilaç-eşleşen ilaç ikilileri sırasıyla matched_drugs_with_drugbank_id_substring_dict_4 ve matched_drugs_with_drugbank_id_substring_matches_dict_4 sözlüklerine aktarıldı.

matched_drugs_with_drugbank_id_substring_dict_4 içindeki boş değerli anahtarlar silindi ve matched_drugs_with_drugbank_id_substring_dict_4_nan_removed sözlüğü oluşturuldu.

common_drugbank_id_substring_matches_dict_3_edited sözlüğü, common_drugbank_id_substring_matches_dict_3 sözlüğü üzerindeki anlamsız eşleşmeler çıkarılarak oluşturuldu. common_drugbank_id_substring_matches_dict_3_edited ve merged_df_drugs_without_drugbank_id_list değişkenleri kullanılarak drug_table_3_name_drugbank_id_dict üzerinde tam karşılık araması yapılıp ilaç-drugbank kimliği ve ilaç-eşleşen ilaç ikilileri sırasıyla matched_drugs_with_drugbank_id_substring_dict_5 ve matched_drugs_with_drugbank_id_substring_matches_dict_5 sözlüklerine aktarıldı.

matched_drugs_with_drugbank_id_substring_dict_5 içindeki boş değerli anahtarlar silindi ve matched_drugs_with_drugbank_id_substring_dict_5_nan_removed sözlüğü oluşturuldu.

matched_drugs_with_drugbank_id_substring_dict_1_nan_removed sözlüğü diğer matched_drugs_with_drugbank_id_substring_dict_2,

matched_drugs_with_drugbank_id_substring_dict_3_nan_removed, matched_drugs_with_drugbank_id_substring_dict_4_nan_removed, matched_drugs_with_drugbank_id_substring_dict_5_nan_removed sözlükleri ile güncellenmiştir.

**Tablo 8.5.** PubChem kimliği alt metin araması eşleşme sözlükleri ve bu sözlüklerin ilaç-PubChem kimliği içeren karşılıkları.

| **Pubchem kimliği alt metin araması eşleşme sözlükleri** | **İlgili ilaç-PubChem kimliği içeren sözlükler** |
|---|---|
| pubchem_id_substring_matches_dict_1 | df3_name_pubchem_id_dict |
| pubchem_id_substring_matches_dict_2 | gdsc_synonym_info_name_pubchem_combined_dict |
| common_pubchem_id_substring_matches_dict_1 | drug_table_1_gdsc_name_pubchem_id_dict |
| common_pubchem_id_substring_matches_dict_2 | drug_table_2_name_pubchem_id_dict |
| common_pubchem_id_substring_matches_dict_3 | drug_table_3_name_pubchem_id_dict |

merged_df_drugs_without_pubchem_id_list_from_df tablosu mapped_pubchem_pubchem_smiles_df tablosu içindeki ‘pubchem_id' sütununda boş olan satırlar ile oluşturuldu. merged_df_drugs_without_pubchem_id_list listesi ise bu yeni tablodaki DRUG_NAME isimli sütundaki özgün ilaç isimleri ile hazırlandı.

pubchem_id_substring_matches_dict_1_edited sözlüğü, pubchem_id_substring_matches_dict_1 sözlüğü üzerindeki anlamsız eşleşmeler çıkarılarak oluşturuldu. pubchem_id_substring_matches_dict_1_edited ve merged_df_drugs_without_pubchem_id_list değişkenleri kullanılarak df3_name_pubchem_id_dict üzerinde tam karşılık araması yapılıp ilaç-PubChem kimliği ve ilaç-eşleşen ilaç ikilileri sırasıyla matched_drugs_with_pubchem_id_substring_dict_1 ve matched_drugs_with_pubchem_id_substring_matches_dict_1 sözlüklerine aktarıldı.

matched_drugs_with_pubchem_id_substring_dict_1 içindeki boş değerli anahtarlar silindi ve

matched_drugs_with_pubchem_id_substring_dict_1_nan_removed sözlüğü oluşturuldu.

pubchem_id_substring_matches_dict_2 ve merged_df_drugs_without_pubchem_id_list değişkenleri kullanılarak gdsc_synonym_info_name_pubchem_combined_dict üzerinde tam karşılık araması yapılıp ilaç-PubChem kimliği ve ilaç-eşleşen ilaç ikilileri sırasıyla matched_drugs_with_pubchem_id_substring_dict_2 ve matched_drugs_with_pubchem_id_substring_matches_dict_2 sözlüklerine aktarıldı.

matched_drugs_with_pubchem_id_substring_dict_2 içindeki boş değerli anahtarlar silindi ve matched_drugs_with_pubchem_id_substring_dict_2_nan_removed sözlüğü oluşturuldu.

common_pubchem_id_substring_matches_dict_1 ve merged_df_drugs_without_pubchem_id_list değişkenleri kullanılarak drug_table_1_gdsc_name_pubchem_id_dict üzerinde tam karşılık araması yapılıp ilaç-PubChem kimliği ve ilaç-eşleşen ilaç ikilileri sırasıyla matched_drugs_with_pubchem_id_substring_dict_3 ve matched_drugs_with_pubchem_id_substring_matches_dict_3 sözlüklerine aktarıldı.

matched_drugs_with_pubchem_id_substring_dict_3 içindeki boş değerli anahtarlar silindi ve matched_drugs_with_pubchem_id_substring_dict_3_nan_removed sözlüğü oluşturuldu.

common_pubchem_id_substring_matches_dict_2 ve merged_df_drugs_without_pubchem_id_list değişkenleri kullanılarak drug_table_2_name_pubchem_id_dict üzerinde tam karşılık araması yapılıp ilaç-PubChem kimliği ve ilaç-eşleşen ilaç ikilileri sırasıyla

matched_drugs_with_pubchem_id_substring_dict_4 ve matched_drugs_with_pubchem_id_substring_matches_dict_4 sözlüklerine aktarıldı.

matched_drugs_with_pubchem_id_substring_dict_4 içindeki boş değerli anahtarlar silindi ve matched_drugs_with_pubchem_id_substring_dict_4_nan_removed sözlüğü oluşturuldu.

common_pubchem_id_substring_matches_dict_3_edited sözlüğü, common_pubchem_id_substring_matches_dict_3 sözlüğü üzerindeki anlamsız eşleşmeler çıkarılarak oluşturuldu. pubchem_id_substring_matches_dict_1_edited ve merged_df_drugs_without_pubchem_id_list değişkenleri kullanılarak drug_table_3_name_pubchem_id_dict üzerinde tam karşılık araması yapılıp ilaç-PubChem kimliği ve ilaç-eşleşen ilaç ikilileri sırasıyla matched_drugs_with_pubchem_id_substring_dict_5 ve matched_drugs_with_pubchem_id_substring_matches_dict_5 sözlüklerine aktarıldı.

matched_drugs_with_pubchem_id_substring_dict_5 içindeki boş değerli anahtarlar silindi ve matched_drugs_with_pubchem_id_substring_dict_5_nan_removed sözlüğü oluşturuldu.

matched_drugs_with_pubchem_id_substring_dict_1_nan_removed sözlüğü matched_drugs_with_pubchem_id_substring_dict_2_nan_removed, matched_drugs_with_pubchem_id_substring_dict_3_nan_removed, matched_drugs_with_pubchem_id_substring_dict_4_nan_removed, matched_drugs_with_pubchem_id_substring_dict_5_nan_removed sözlükleriyle güncellendi. Bu sözlükten 'Ulixertinib' elemanı boş değer içerdiği için silindi.

**Tablo 8.6.** SMILES gösterimi alt metin araması eşleşme sözlükleri ve bu sözlüklerin ilaç-SMILES eşleşmesi içeren karşılıkları.

| SMILES gösterimi alt metin araması eşleşme sözlükleri | İlgili ilaç-SMILES eşleşmesi içeren sözlükler |
|---|---|
| common_smiles_substring_matches_dict_1 | drug_table_1_gdsc_name_smiles_dict |
| common_smiles_substring_matches_dict_2 | drug_table_2_name_smiles_dict |
| common_smiles_substring_matches_dict_3 | drug_table_3_name_smiles_dict |
| common_smiles_substring_matches_dict_4 | df3_name_smiles_dict |

merged_df_drugs_without_smiles_list_from_df tablosu mapped_drugbank_pubchem_smiles_df tablosu içindeki 'SMILES' sütununda boş olan satırlar ile oluşturuldu. merged_df_drugs_without_smiles_list listesi ise bu yeni tablodaki DRUG_NAME isimli sütundaki özgün ilaç isimleri ile hazırlandı.

common_smiles_substring_matches_dict_1 ve merged_df_drugs_without_smiles_list değişkenleri kullanılarak drug_table_1_gdsc_name_smiles_dict üzerinde tam karşılık araması yapılıp ilaç-SMILES eşleşmesi matched_drugs_with_smiles_substring_dict_1 sözlüğüne aktarıldı. matched_drugs_with_smiles_substring_dict_1 içindeki boş değerli anahtarlar silindi ve matched_drugs_with_smiles_substring_dict_1_nan_removed sözlüğü oluşturuldu.

common_smiles_substring_matches_dict_2 ve merged_df_drugs_without_smiles_list değişkenleri kullanılarak drug_table_2_name_smiles_dict üzerinde tam karşılık araması yapılıp ilaç-SMILES eşleşmesi matched_drugs_with_smiles_substring_dict_2 sözlüğüne aktarıldı. matched_drugs_with_smiles_substring_dict_2 içindeki boş değerli anahtarlar silindi ve matched_drugs_with_smiles_substring_dict_2_nan_removed sözlüğü oluşturuldu.

common_smiles_substring_matches_dict_3 ve merged_df_drugs_without_smiles_list değişkenleri kullanılarak drug_table_3_name_smiles_dict üzerinde tam karşılık araması yapılıp ilaç-SMILES

eşleşmesi matched_drugs_with_smiles_substring_dict_3 sözlüğüne aktarıldı. matched_drugs_with_smiles_substring_dict_3 içindeki boş değerli anahtarlar silindi ve atched_drugs_with_smiles_substring_dict_3_nan_removed sözlüğü oluşturuldu.

common_smiles_substring_matches_dict_4 ve merged_df_drugs_without_smiles_list değişkenleri kullanılarak df3_name_smiles_dict üzerinde tam karşılık araması yapılıp ilaç-SMILES eşleşmesi matched_drugs_with_smiles_substring_dict_4 sözlüğüne aktarıldı.

matched_drugs_with_smiles_substring_dict_4 içindeki boş değerli anahtarlar silindi ve matched_drugs_with_smiles_substring_dict_4_nan_removed sözlüğü oluşturuldu.

matched_drugs_with_smiles_substring_dict_3_nan_removed sözlüğünden 'FTY-720' anahtarı anlamsız eşleşme nedeniyle silindi.

matched_drugs_with_smiles_substring_dict_1_nan_removed sözlüğü matched_drugs_with_smiles_substring_dict_2_nan_removed, matched_drugs_with_smiles_substring_dict_3_nan_removed, matched_drugs_with_smiles_substring_dict_4_nan_removed sözlükleri ile güncellendi.

'gdsc.smi' isimli dosya df6_gdsc_smiles_df değişkenine aktarıldı. Tablounun ilk sütunu SMILES; ikinci sütunu drug_name isimleriyle değiştirildi.

df6_gdsc_smiles_df_drug_name_smiles_dict sözlüğü df6_gdsc_smiles_df üzerindeki name ve drug_name sütunları kullanılarak hazırlandı.

df6_gdsc_smiles_df_drug_name_smiles_dict ve merged_df_drugs_without_smiles_list değişkenleri kullanılarak tam karşılık araması yapıldı. Ve, ilaç-SMILES eşleşmeleri df6_matched_drug_name_smiles_dict sözlüğüne atandı. Alt metin araması yapılarak da karşılıklar ve eşleşmeler bulunup sırasıyla df6_matched_drug_name_smiles_substring_dict ve df6_matched_drug_name_smiles_substring_matches_dict sözlüklerine aktarıldı.

‘LINCS.csv’ adlı dosya df7_lincs_smiles_df değişkenine aktarıldı.

df7_lincs_pert_iname_smiles_dict sözlüğü df7_lincs_smiles_df üzerindeki pert_iname ve canonical_smiles sütunları kullanılarak hazırlandı.

df7_lincs_pert_iname_smiles_dict ve merged_df_drugs_without_smiles_list değişkenleri kullanılarak tam karşılık araması yapıldı. Ve, ilaç-SMILES eşleşmeleri df6_matched_drug_name_smiles_dict sözlüğüne atandı. Alt metin araması yapılarak da karşılıklar ve eşleşmeler bulunup sırasıyla df7_matched_drug_name_smiles_substring_dict ve df7_matched_drug_name_smiles_substring_matches_dict sözlüklerine aktarıldı.

'drug_info_candle.txt' adlı dosya df8_drug_info_df değişkenine aktarıldı.

Tablodaki 'PUBCHEM' sütunu tam sayı olarak yeniden aynı sütuna atandı. Boş değerli df8_pubchem_id_dict sözlüğü oluşturuldu ve tablo üzerindeki ilk sütundaki ‘GDSC’ içeren satırlar bulunup ilgili ilaç ismi ve PubChem kimliği çifti bu sözlüğe yazıldı. Boş değer içeren anahatarlar keys_with_nan listesine alındı ve bu elemanlar sözlükten silindi.

df8_drug_name_smiles_dict boş sözlüğü oluşturularak df8_drug_info_df tablosundaki ilk sütunda ‘GDSC’ içeren satırlar özelinde ilaç-SMILES çiftleri bulunup sözlüğe aktarıldı. df8_drug_name_smiles_dict ve merged_df_drugs_without_smiles_list değişkenleri kullanılarak ilaç-SMILES çiftlerini bulmak için tam karşılık araması yapıldı ve sonuçlar df8_matched_drug_name_smiles_dict sözlüğüne yazıldı. Bu sözlükteki boş değer içeren elemanlar göz ardı edilerek df8_matched_drug_name_smiles_dict_nan_removed sözlüğü oluşturuldu.

df8_drug_name_smiles_dict ve merged_df_drugs_without_smiles_list değişkenleri kullanılarak ilaç-SMILES çiftlerini bulmak için alt metin araması yapıldı. Karşılık ve eşleşme sonuçları sırasıyla df8_matched_drug_name_smiles_substring_dict, df8_matched_drug_name_smiles_substring_matches_dict sözlüklerine yazıldı.

df8_matched_drug_name_smiles_substring_dict içindeki boş değer içeren elemanlar göz ardı edilerek df8_matched_drug_name_smiles_substring_dict_nan_removed sözlüğü oluşturuldu.

'structure links_DrugBank.csv' adlı dosya df9_structure_info_df değişkenine atandı. Tablo üzerindeki sütun isimleri 'DrugBank ID': 'DrugBank_ID', 'PubChem Compound ID': 'PubChem_Compound_ID', 'ChEBI ID':' ChEBI_ID', 'ChEMBL ID': 'ChEMBL_ID' şeklinde değilştirildi.

'PubChem_Compound_ID' ve 'ChEBI_ID' sütunlarındaki değerler tam sayı olarak değiştirildi.

df9_name_drugbank_id_dict, df9_name_pubchem_id_dict, df9_name_chebi_id_dict, df9_name_chembl_id_dict ve df9_name_smiles_dict sözlükleri Name ve sırasıyla DrugBanK_ID, PubChem_Compound_ID, ChEBI_ID, ChEMBL_ID, SMILES sütunları kullanılarak hazırlandı.

drugbank_id_dict_1 sözlüğü matched_drugs_with_drugbank_id_substring_dict_1_nan_removed ile güncellendi.

pubchem_id_dict_1_nan_removed sözlüğü sırasıyla matched_drugs_with_pubchem_id_substring_dict_1_nan_removed ve df8_pubchem_id_dict ile güncellendi.

df6_matched_drug_name_smiles_dict sözlüğü sırasıyla matched_drugs_with_smiles_substring_dict_1_nan_removed, df7_matched_drug_name_smiles_dict, df8_matched_drug_name_smiles_dict_nan_removed, df9_name_smiles_dict ile güncellendi.

common_smiles_dict_1_nan_removed sözlüğü ise df6_matched_drug_name_smiles_dict ile güncellendi.

mapped_drugbank_pubchem_smiles_df üzerinde 'drugbank_2', 'pubchem_id_2', 'SMILES_2' sütunları DRUG_NAME sütunu baz alınarak sırasıyla drugbank_id_dict_1, pubchem_id_dict_1_nan_removed, common_smiles_dict_1_nan_removed sözlükleri ile oluşturuldu.

**Tablo 8.7.** İlaçlara ait kimlik numaraları ve SMILES görünümlerini almak için kullanılan diğer dosyalar ile ilgili bilgiler.

| Dosya adı | Erişim adresi | Dosyadaki tablo boyutu |
|---|---|---|
| DrugBank_5.1.6_all_structure_links_SMILES_InChI_Xref.csv | - | 11405, 17 |
| drugbank vocabulary.csv | https://go.drugbank.com/releases/5-1-7/downloads/all-drugbank-vocabulary | 13580, 7 |
| Drug_listFri Nov 27 09_00_17 2020_edited.csv | https://www.cancerrxgene.org/compounds | 565, 9 |
| drug_annotation_table-2020-11-27T10 40 48-0500.csv | https://pharmacodb.pmgenomics.ca/download?drug_annotation=y | 759, 15 |
| drugs.tsv | https://www.pharmgkb.org/downloads | 3429, 25 |
| chemicals.tsv | https://www.pharmgkb.org/downloads | 4170, 24 |
| gdsc.smi | https://ibm.ent.box.com/v/paccmann-pytoda-data/file/548614907232 | 209, 2 |
| LINCS.csv | https://raw.githubusercontent.com/bhklab/DNF/alex-changes/Data/LINCS.csv | 20326, 28 |
| drug_info_candle.txt | https://ftp.mcs.anl.gov/pub/candle/public/benchmarks/Pilot1/combo/drug_info | 846, 6 |

mapped_drugbank_pubchem_smiles_df üzerindeki sütun isimleri 'drugbank_2': 'DrugBank_ID', 'pubchem_id_2': 'PubChem_Compound_ID', 'chembl_id_2': 'ChEMBL_ID', 'chebi_id_2':'ChEBI_ID' şeklinde değiştirildi.

mapped_drugbank_pubchem_smiles_df_v3 değişkeni üzerine mapped_drugbank_pubchem_smiles_df tablosundaki sütunlar 'DRUG_NAME', 'SMILES', 'CELL_LINE_NAME', 'DrugBank_ID', 'PubChem_Compound_ID', 'ChEMBL_ID', 'ChEBI_ID' sırasıyla aktarıldı.

mapped_drugbank_pubchem_smiles_df_v3 tablosu ‘mapped_drugbank_pubchem_smiles_df_v3.txt’ dosya üzerine yazıldı.

**İlaçların SMILES Dizisine Sahip Olanları İçin Ilaç Parmak İzi Değerlerinin Oluşturulması**

Pandas ve NumPy kütüphaneleri hazırlandıktan sonra 449 ilaç ve SMILES dizilerinin bulunduğu "GDSC_drug_smiles_df_v1.txt" dosyasındaki tablo bir değişkene atanmıştır.

Sys ve RDkit kütüphaneleri kullanılmaya uygun hale getirilip tabloya ECFP4 isimli boş değerli sütun eklenmiştir.

İlaçların SMILES dizilerinin bulunduğu sütunldaki değerlerin teker teker işlenerek ECFP4 sütununa karşılıklarının yazıldığı bir döngü oluşturulmuştur. Bu döngüde, SMILES dizileri GetMorganFingerprintAsBitVect metodu ve radius (2 değeri atanmıştır), nBits (1024 değeri atanmıştır) parametreleri yardımıyla ilaç parmak izlerine dönüştürülmüştür. Parmak izi oluşturma işlemi iki adımda gerçekleştiği ve nBits parametresi 1024 olarak belirlendiği için parmak izleri, ECFP_4 formatında ve 1024 elemanlı (bit) olarak üretilmiştir.

Tabloda yer alan boş değerli satırlara metin olarak “NaN” değeri atanmıştır. Sonuç tablosu bir dosyaya yazdırılıp kaydedilmiştir.

### 8.2.3. Vektör Matrisinin Oluşturulması

#### 35542 Uzunluklu Vektörler İçeren Matrisin Oluşturulması

Pandas ve NumPy kütüphaneleri hazır hale getirilip GDSC_gene_exp_extended_genes_cell_lines_v2.txt, GDSC_mutation_extended_genes_cell_lines_v3.txt, GDSC_methylation_extended_genes_cell_lines_v2.txt, GDSC_cnv_picnic_extended_genes_cell_lines_v3.txt farklı birer değişkene atanarak sırasıyla gen ifade, mutasyon, metilasyon, KSD veri tiplerine ait tablolar oluşturulmuştur.

Gen ifade verisi için, gen temelli olarak sütunlardaki değerlerin ortalamaları alınarak bir tablo oluşturulmuştur. Bu tablonun boş değerli satırlarına metin değeri olarak “NaN” atanmıştır. İki sütunlu bu tablodan bir sözlük üretilip kaydedilmiştir.

Gen ifade verisi için satır bazında doluluk oranının hesaplandığı bir tablo hazırlanmıştır. Gen ismi sütunundan sonra toplam dolu sütun sayısı ve doluluk yüzdesi sütunu gelmektedir. Tablonun oluşturulmasında gen ifade ortalama sözlüğünden yararlanılıp boş değerli satırlarda dolu sütun sayısı için “NaN”, yüzdesi için ise 0 değeri girilmiştir. Bu tablo üzerinde filtreleme işlemi yapılıp yüzde 90 ve üzeri doluluk oranına sahip olan genler bırakılmıştır.

KSD verisi için, gen temelli olarak sütunlardaki değerlerin ortalamaları alınarak bir tablo oluşturulmuştur. Bu tablonun boş değerli satırlarına metin değeri olarak “NaN” atanmıştır. İki sütunlu bu tablodan bir sözlük üretilip kaydedilmiştir.

KSD verisi için satır bazında doluluk oranının hesaplandığı bir tablo hazırlanmıştır. Gen ismi sütunundan sonra toplam dolu sütun sayısı ve doluluk yüzdesi sütunu gelmektedir. Tablonun oluşturulmasında KSD ortalama sözlüğünden yararlanılıp boş değerli satırlarda dolu sütun sayısı için “NaN”, yüzdesi için ise 0 değeri girilmiştir. Bu tablo üzerinde filtreleme işlemi yapılıp yüzde 87 ve üzeri doluluk oranına sahip olan genler bırakılmıştır.

En büyük gen listelerine sahip olan gen ifade ve KSD'nin filtrelenmiş olan tablolarında bulunan genlerin ortak kümesi alınarak bir listeye atanmıştır. Bu liste kullanılarak tüm omik veri tiplerine ait ana tablolar gen isimleri sütunundan filtrelenmiştir.

Mutasyon verisi için filtrelenen ana tablodan, gen temelli olarak sütunlarda 1 değeri içerenler dikkate alınarak bir tablo oluşturulmuştur. Tablo, gen ismi ve gen satırında bulunan toplam 1 değeri sayısını belirten sütundan oluşmaktadır. Bu tablo üzerinde yeni bir sütun yaratılarak gen temelli doluluk yüzdeleri eklenmiştir. Yüzde 5 ve üzeri eşik değeri uygulanarak filtreleme yapılmıştır. Ayrıca kalan gen isimleri bir listeye atanmıştır.

Metilasyon verisi için filtrelenen ana tablodan, gen temelli olarak sütunlarda 1 değeri içerenler dikkate alınarak bir tablo oluşturulmuştur. Tablo, gen ismi ve gen satırında bulunan toplam 1 değeri sayısını belirten sütundan oluşmaktadır. Bu tablo üzerinde yeni bir sütun yaratılarak gen temelli doluluk yüzdeleri eklenmiştir. Yüzde 1 ve üzeri eşik değeri uygulanarak filtreleme yapılmıştır. Ayrıca kalan gen isimleri bir listeye atanmıştır.

Mutasyon ve metilasyon için oluşturulan gen listeleri kullanılarak filtrelenmiş ana tablolar üzerinden bu genlere ait satırlar elde edilip oluşan tablolar birer değişkene atanmıştır.

Tüm omik veri tiplerinin ortak gen ve hücre hattı isimlerinin bulunduğu final_gene_cell_line_lists_combined.txt dosyasındaki tablo bir değişkene atanmıştır. Gen ve hücre hatlarının bulunduğu sütunlar ise birer listeye alınıp kaydedilmiştir. Hücre hattına ait listeden boş değerli elemanlar silinmiştir.

Tüm omik veri tablolarında "Capan-2","CAPAN-2"; "DiFi", "DIFI"; "HuTu-80","HUTU-80";"Jurkat", "JURKAT"; "MMAc-SF", "MMAC-SF" gibi aynı hücre hattına ait olan sütunlar yer almaktadır. Tüm tablolarda sütun ismi değişikliğine gidilerek, bu

isimlerin büyük harflerle yazılmış olanları küçük harf içeren asıllarıyla değiştirilmiştir. Ek olarak, ortak hücre hatlarının bulunduğu listeden "Jurkat", "JURKAT", "CAPAN-2", "DIFI", "HUTU-80", "MMAC-SF" isimli olanlar değerlendirme dışı bırakılmıştır. İki ayrı sütunda değeri bulunan hücre hatları ("Capan-2", "DiFi", "HuTu-80", "MMAc-SF") için ayrı şekilde birleştirim yapılacağı için, bu isimleri barındıran farklı bir liste oluşturulmuştur.

Tek sütunda değeri bulunan ortak hücre hatlarına ait liste kullanılarak ilk vektör matrisi oluşturulmuştur. Bunun için, değerlendirilen her hücre hattı için aynı prosedür uygulanmıştır. Öncelikle, hücre hattına ait sütunlar omik veri tablolarından gen ifade ve KSD sırasıyla alınıp düşey eksende birleştirilmiştir. Oluşan tabloda, boşluk değerli hücrelere “NaN” değeri atanmıştır. Gen ifade ve KSD sütunlarındaki değerler birer listeye alınmıştır. Bu listeler üzerinde önceden oluşturulan sözlükler yardımıyla “NaN” içerikli bölümlere o gene ait ortalama değer atanmıştır. Mutasyon ve metilasyon verileri ilgili tablolardan alınıp ayrı listelere kaydedilmiştir. Son olarak, gen ifade, mutasyon, metilasyon ve KSD listeleri bu sırayla birleştirilip bir listeye aktarılmıştır. Düzenlenen hücre hattına ait dosya ismi düzenlendikten sonra, liste bir tablo şeklinde bir dosyaya yazdırılıp kaydedilmiştir. Aynı prosedür iki ayrı sütunda değeri bulunan hücre hatlarına da uygulanmıştır.

**1910 Uzunluklu Vektörler İçeren Matrisin Oluşturulması**

Yukarıda 35542 uzunluğundaki dosyaların hazırlanma aşamaları, uygulanan prosedür 1910 uzunluklu vektör için de uygulanmıştır. Ancak, vektör uzunluğunu daha da kısaltmak için ek işlemler yapılmıştır.

Öncelikle, L1000_gene_list.txt dosyasından L1000 genleri bir listeye aktarılmıştır. Omik veri tablolarının son halindeki gen isimleri ayrı birer listeye alınmıştır. Her omik verisine ait gen listelerinin L1000 listesi ile ortaklıkları alınıp ayrı listelere aktarılmıştır.

GDSC_986_cell_lines_matched_with_TCGA_tissue_names.txt dosyasından elde edilen doku ismi bulunan 986 hücre hattına ait isimler bir listeye aktarılmıştır.

Her omik veri tablosu için aşağıdaki prosedür uygulanıp hepsi için yeni birer tablo oluşturulmuştur. L1000 ortak gen listeleri ile filtrelenen ana tablo üzerinden yeniden eleme yapılarak ilgili satırlar elde edilmiştir. Oluşan tablodan 986 hücre hattına ait sütunlar filtrelenmiştir. Bu tabloya, gen isimleri sütunu eklendikten sonra sindirim sistemine ait hücre hatları bu tablodan elde edilmiştir. Transpoze edilen bu tablo bir değişkene atanmıştır. Son adımda yalnızca gen ifade ve KSD tablolarındaki boşluk içeren yerler gen temelli olarak ortalama değerle doldurulmuştur. Sonuç tablolarının içerdikleri gen sayıları şu şekildedir; gen ifade (896), mutasyon (107), metilasyon (11), KSD (11).

Omik verilere ait son tabloların sütun değerleri birer sayı atanarak sözlükler oluşturulmuştur. Bu sözlükler yardımıyla tabloların sütun isimleri değiştirilmiştir.

Son olarak, gen ifade, mutasyon metilasyon, KSD verileri sırasıyla dikey eksende birleştirildikten sonra bir dosyaya yazdırılıp kaydedilmiştir.

**1989 Uzunluklu Vektörler İçeren Matrisin Oluşturulması**

Veri içi alan analizinde kullanılmak üzere 1989 uzunluğa sahip vektörler oluşturulmuştur.

Pandas ve NumPy kütüphaneleri hazır hale getirilip GDSC_gene_exp_extended_genes_cell_lines_v2.txt, GDSC_mutation_extended_genes_cell_lines_v3.txt, GDSC_methylation_extended_genes_cell_lines_v2.txt, GDSC_cnv_picnic_extended_genes_cell_lines_v3.txt farklı birer değişkene atanarak sırasıyla gen ifade, mutasyon, metilasyon, KSD veri tiplerine ait tablolar oluşturulmuştur.

Gen ifade verisi için, gen temelli olarak sütunlardaki değerlerin ortalamaları alınarak bir tablo oluşturulmuştur. Bu tablonun boş değerli satırlarına metin değeri olarak “NaN” atanmıştır. İki sütunlu bu tablodan bir sözlük üretilip kaydedilmiştir.

Gen ifade verisi için satır bazında doluluk oranının hesaplandığı bir tablo hazırlanmıştır. Gen ismi sütunundan sonra toplam dolu sütun sayısı ve doluluk yüzdesi sütunu gelmektedir. Tablonun oluşturulmasında gen ifade ortalama sözlüğünden yararlanılıp boş değerli satırlarda dolu sütun sayısı için “NaN”, yüzdesi için ise 0 değeri girilmiştir. Bu tablo üzerinde filtreleme işlemi yapılıp yüzde 90 ve üzeri doluluk oranına sahip olan genler bırakılmıştır.

KSD verisi için, gen temelli olarak sütunlardaki değerlerin ortalamaları alınarak bir tablo oluşturulmuştur. Bu tablonun boş değerli satırlarına metin değeri olarak “NaN” atanmıştır. İki sütunlu bu tablodan bir sözlük üretilip kaydedilmiştir.

KSD verisi için satır bazında doluluk oranının hesaplandığı bir tablo hazırlanmıştır. Gen ismi sütunundan sonra toplam dolu sütun sayısı ve doluluk yüzdesi sütunu gelmektedir. Tablonun oluşturulmasında KSD ortalama sözlüğünden yararlanılıp boş değerli satırlarda dolu sütun sayısı için “NaN”, yüzdesi için ise 0 değeri girilmiştir. Bu tablo üzerinde filtreleme işlemi yapılıp yüzde 87 ve üzeri doluluk oranına sahip olan genler bırakılmıştır.

En büyük gen listelerine sahip olan gen ifade ve KSD’nin filtrelenmiş olan tablolarında bulunan genlerin ortak kümesi alınarak bir listeye atanmıştır. Bu liste kullanılarak tüm omik veri tiplerine ait ana tablolar gen isimleri sütunundan filtrelenmiştir.

Mutasyon verisi için filtrelenen ana tablodan, gen temelli olarak sütunlarda 1 değeri içerenler dikkate alınarak bir tablo oluşturulmuştur. Tablo, gen ismi ve gen satırında bulunan toplam 1 değeri sayısını belirten sütundan oluşmaktadır. Bu tablo üzerinde yeni bir sütun yaratılarak gen temelli doluluk yüzdeleri eklenmiştir. Yüzde 5 ve üzeri eşik değeri uygulanarak filtreleme yapılmıştır. Ayrıca kalan gen isimleri bir listeye atanmıştır.

Metilasyon verisi için filtrelenen ana tablodan, gen temelli olarak sütunlarda 1 değeri içerenler dikkate alınarak bir tablo oluşturulmuştur. Tablo, gen ismi ve gen satırında bulunan toplam 1 değeri sayısını belirten sütundan oluşmaktadır. Bu tablo üzerinde yeni bir sütun yaratılarak gen temelli doluluk yüzdeleri eklenmiştir. Yüzde 1 ve üzeri eşik değeri uygulanarak filtreleme yapılmıştır. Ayrıca kalan gen isimleri bir listeye atanmıştır.

L1000_gene_list.txt dosyasındaki tablo bir değişkene atanarak kaydedilmiştir. Ardından, tablodaki pr_gene_symbol sütununda bulunan gen isimleri bir listeye aktarılmıştır. Yukarıda filtrelenen omik veri tablolarından elde edilen gen ismi listelerinin ayrı ayrı olarak L1000 gen listesi ile ortaklıkları bulunup listelere atanmıştır. Oluşan bu listeler ile ana omik veri tablolarında gen ismi sütunu üzerinden filtreleme yapılmıştır.

GDSC_986_cell_lines_matched_with_TCGA_tissue_names.txt dosyasındaki 986 hücre hattına ait bilgilerin bulunduğu tablo, bir değişkene atanmıştır. Ardından, tablo doku ismi temelli olarak gruplanmıştır. Bir sözlük oluşturularak doku isimleri anahtar; dokuda bulunan hücre hatları liste olarak değer değişkeni şeklinde atanmıştır. Ayrıca, tablodaki hücre hatları isimleri bir listede kaydedilmiştir.

Filtrelenen omik veri tablolarından belirli hücre hatları iki ayrı sütunda değerleri görüldüğü için çıkarılmıştır. Gen ifade için "CAPAN-2", "DiFi", "HuTu-80", "MMAc-SF"; Mutasyon için "Capan-2", "DIFI", "HUTU-80", "MMAC-SF"; Metilasyon için "Capan-2", "DIFI", "HUTU-80", "MMAC-SF"; KSD için ise "Capan-2", "DIFI", "HUTU-80", "MMAC-SF" isimli sütunlar çıkarılmıştır. Devamında ise, bazı hücre hatlarının isimleri değiştirilmiştir. Gen ifade için değişim sözlüğü şu şekilde oluşturulmuştur, "DIFI": "DiFi", "HUTU-80": "HuTu-80", "MMAC-SF": "MMAc-SF". Mutasyon, metilasyon ve KSD için değişim sözlüğü şu şekilde oluşturulmuştur, "CAPAN-2": "Capan-2".

Elde edilen filtrelenmiş omik veri tablolarının boyutları şu şekildedir, gen ifade (923, 1123), mutasyon (110, 1123), metilasyon (12, 1123), KSD (944, 1123). Bu

tablolardaki gen isimleri birer dosyaya yazdırılıp kaydedilmiştir. Tabloların indeks kısımları yeniden 0'dan başlatılmak üzere düzenlenmiştir.

Ana tablolarda ikili sütun bulunduran hücre hatları (3 adet) ile tek sütun bulunduran hücre hatları (983 adet) için ayrı aşamalarda aynı prosedür izlenerek vektör matrisleri oluşturulmuştur. Uygulanan prosedür aşağıda belirtildiği gibidir.

Öncelikle, incelemeye dahil edilen hücre hattı sayısı kadar satır ve 1989 (bir vektör uzunluğu) sütun içeren boş değerli bir tablo oluşturulmuştur. Ardından, bir döngü içinde her hücre hattının ayrı ayrı değerlendirildiği bir akış içinde, gen ifade ve KSD tablolarından ilgili hücre hattına ait sütunlar bir tabloda dikey eksenli birleştirilmiştir. Bu tablodaki boş değerli yerlere "NaN" değeri atanmıştır. Tablodaki gen ifade ve KSD değerleri ayrı listelere atanmıştır. Bu listeler ayrı döngüler içinde değerlendirilip "NaN" değerli elemanları için önceden oluşturulan ortalama (gen ifade için) veya medyan (KSD için) değerler atanmıştır. Mutasyon ve metilasyon tablolarından da ilgili hücre hattının değerleri birer listeye aktarılmıştır. Boş bir listede, sırasıyla gen ifade, mutasyon, metilasyon, KSD listelerine ait değerler birleştirilmiştir. Hücre hattının vektörü olarak bu liste boş tablonun satırlarına sırasıyla yerleştirilmiştir. Oluşturulan tablonun ilk sütununa denk gelecek şekilde hücre hattı isimleri yazdırılmıştır. Sonuç olarak elde edilen iki vektör matrisi yatay eksende birleştirilip indekste yer alan hücre hatları alfabetik olarak sıralanmıştır.

Sonuç tablosu bir dosyaya yazdırılarak kaydedilmiştir. Ek olarak, sonuç tablosu üzerinde yapılan işlemle, her dokunun hücre hatları tablodan filtrelenip o dokuya özel dosya olarak kaydedilmiştir.

**Hücre Hattı Özellik Vektörlerinde Kullanılacak Gen Sayısının Azaltılması**

Modellenecek verinin ana yapısını oluşturan hücre hattı özellik vektörlerinin uzunluğu hem tahmin performansını hem de model eğitiminin tamamlanacağı süreyi etkilemektedir. Bu nedenle, 35542 ve 1910 vektör uzunluklu sindirim sistemi verisi

üzerinden farklı hiperparametre ve değerleri için tahmin modelleri oluşturulup birbirleriyle karşılaştırılmıştır. Her iki vektör tipi için de aynı işlem süreçleri uygulandığından, sadece 35542 vektör uzunluklu matris üzerinden yapılan aşamalar sunulmuştur.

Pandas, NumPy, Math kütüphaneleri hazır hale getirildikten sonra, GDSC_drug_response_df_cell_line_fingerprint_pIC50s_for_rforest1.txt ilaç yanıtı verisini içeren dosyadaki tablo bir değişkene atanmıştır.

Sklearn kütüphanesinin preprocessing (StandardScaler), model_selection (train_test_split, GridSearchCV, cross_val_score), metrics, ensemble (RandomForestRegressor) metotları hazır hale getirilmiştir.

digestive_system_tissue_based_combined_cell_line_features_35K_rows.txt isimli GDSC sindirim sistemi dosyasındaki tablo bir değişkene atanmıştır. Bu tabloda bulunan hücre hatları bir listeye aktarılmıştır.

İlaç yanıtı tablosundaki hücre hattı sütunundan sindirim sistemi dokusunda bulunan hücre hatlarına ait satırlar filtrelenmiştir. Ardından, yine bu tablodaki parmak izi sütunundaki değerler bir listeye aktarılmıştır.

Öncelikle, her elemanı birbirinden ayrılan parmak izlerinin depolanacağı boş bir liste oluşturulmuştur. Bir döngü içinde yukarıda oluşturulan parmak izi listesindeki her örnek parçalara ayrılıp bir liste olarak boş liste içine sırasıyla eklenmiştir. Bu yeni listedeki veri bir tablo haline getirilmiştir. Oluşan yeni tabloya hücre hattı isimleri ve ilaç yanıtı değerleri ana tabloda aynı sıra gözetilerek eklenmiştir. Bu tablo vektör matrisinin sağ tarafına ekleneceği için, tablonun tüm sütun isimlerini değiştirmek için bir sözlük yardımıyla her sütun ismi değeri (sayı) vektör tablosunun uzunluğu kadar artırılmıştır.

Vektör matrisi ve ilaç parmak izleri değerlerinin ayrıldığı tablo hücre hattı ismi sütunu üzerinden birleştirilmiştir. Bu yeni tabloda, metin değeri bulunduran sütunlar dışındaki yerlerin hepsi tek duyarlıklı kayan noktalı sayı biçimi (*single precision*) olarak

yeniden yazdırılmıştır. Modelde değişken ve sonuç (etiket) bilgisi olarak kullanılacak veriler farklı değişkenlere aktarılmıştır. Tablo 3.9.'da belirtilen hiperparametreler için belirtilen değerler bir sözlük içinde kaydedilmiştir. RandomForestRegressor metodu yardımıyla değişken ve sonuç değerleri birbirine uydurulup bir objeye atanmıştır. GridSearchCV metodu ile RF objesi, hiperparametre sözlüğü, 5 değeri atanan çapraz doğrulama, MAE skorlama metriği, 4 değeri atanan n_jobs ve 3 değerli verbose metrikleriyle bir objede tutulmuştur. Bu obje ile fit metodu kullanılarak, değişken ve sonuç verisi birbirine uydurulmaya çalışılmıştır. Ardından, cross_val_score metodu ile GridSearchCV objesi, değişken ve sonuç verisi bir arada değerlendirilerek çapraz doğrulama sonuçları bir değişkene atanmıştır.

Yukarıdaki aşamalar Tablo 3.9.'da belirtilen farklı hiperparametre seçenekleriyle denenip MAE skor değeri ve model eğitim süresi kaydedilmiştir (Bkz. Tablo 4.8.).

### 8.2.4. Veri Görselleştirme

Farklı aşamalarda elde edilen değişik uzunluktaki veri tipi ve vektör matrislerinin, gen sayısı azaltma işlemlerinin etkili olduğunun başka bir destekleyicisi olarak t-SNE tekniğiyle görselleştirilmeleri yapılmıştır. Görselleştirmelerde, kaynak tablolar olan omik veri tipleri, 35542 ve 3747 uzunluklarındaki vektörlere sahip matrislerde bulunan omik veri tipleri kullanılmıştır.

**Kaynak Veri**

Pandas, NumPy, Seaborn, Matplotlib, sklearn.manifold (TSNE metodu içerdiği için) kütüphaneleri hazır duruma getirilmiştir. L1000_gene_list.txt dosyasındaki tablo bir değişkene atanmıştır. Tablodaki L1000 gen isimleri bir listeye aktarılmıştır.

GDSC_gene_exp_v2.txt dosyasındaki tablo gen ifade verisine ait bir değişkene aktarılmıştır. Ardından, tablodaki gen isimleri listelenmiştir ve L1000 listesiyle ortaklıkları ayrı bir listede tutulmuştur. Bu işlemler, GDSC_mutation_v1.txt (mutasyon),

GDSC_methylation_v1.txt (metilasyon), GDSC_cnv_picnic_v1.txt (KSD) dosyaları için de yapılmıştır.

GDSC_986_cell_lines_matched_with_TCGA_tissue_names.txt dosyasından hücre hattı ve doku eşleşmelerine ait tablo bir değişkene aktarılmıştır. Bu tablodan, hücre hattı ve doku isimleri birer listede tutulmuştur.

Bu aşamadan sonraki işlemler tüm omik veri tiplerinde aynı şekilde uygulandığı için sadece gen ifade verisi üzerinden anlatım yapılmıştır.

L1000 ortak genleri gen ifade verisi üzerinden filtrelenip bir değişkene atanmıştır. Gen ifade tablosundaki gen isimleri sütunu indeks durumuna getirilip sütun isimleri bir listeye aktarılmıştır. Sütun isimlerindeki hücre hatları ile eşleşmelerin olduğu tablodaki 986 hücre hattının arasındaki ortak isimler bir listeye alınmıştır. Filtrelenen gen ifade verisi üzerinden hücre hatları için ikinci filtreleme bu liste ile yapılmıştır. Oluşan yeni tabloya gen isim sütunu tekrar eklenmiştir. Tablonun görselleştirme algoritması tarafından kullanımını kolaylaştırması için transpozu alınmıştır. Doku listesindeki isimler bir sözlük yardımıyla numaralandırılmıştır. Bu sözlük ile hücre karşılıklarının olduğu tabloya doku numaraları yansıtılmıştır. Bu tablodan hücre hattı isimleri ve doku numaraları sütunlarıyla yeni bir sözlük oluşturulmuştur.

Transpoze edilen tabloda indeks olarak bulunan gen isimleri tablo içine tekrar alınmış ve sütun ismi değiştirilmiştir. Bu tabloya yeni bir sütun olarak doku numaraları son oluşturulan sözlük yardımıyla eklenmiştir. Oluşan tablodan hücre hattı isim sütunu çıkarıldıktan sonra tablodaki boş yerler bulundukları sütundaki ortalama değer ile doldurulmuştur. Tablodaki gen ifade verisi ile alakalı kısımlar değişken; doku numaraları sonuç (etiket) olarak değerlendirilip farklı değişkenlere atanmıştır.

TSNE metodunda kullanılmak üzere şu parametreler ve değerleri seçilip bir obje oluşturulmuştur, n_components (2); verbose (1), random_state (2). Bu değerler iki boyutlu grafiğin tekrar aynı şekilde oluşabilmesi için tercih edilmiştir. Oluşturulan obje ile

fit_transform metodu kullanılıp gen ifade verisi (değişken) için boyut indirgeme işlemi yapılmış ve sonuçlar bir değişkene atanmıştır.

Boş bir tablo oluşturulup doku numaraları bilgisinin olduğu değişken ilk sütun olarak eklenmiştir. Sonrasında, t-SNE grafiğinin ilk komponenti olacak sütuna indirgeme sonuçlarının ilk sütun değerleri; ikinci komponenti için ikinci sütun değerleri aktarılmıştır. Doku isimlerinin numaralandırıldığı sözlüğün ters versiyonu oluşturulmuştur. Bu yeni sözlükle, oluşturduğumuz üç sütunlu tablodaki doku numaraları doku isimlerine geri çevrilmiştir.

Distinctipy kütüphanesi hazır hale getirilmiştir. Get_colors metoduyla rastgele 13 ayrı renk seçilip bir değişkene atanmıştır. Seaborn kütüphanesi metodu olan relplot için parametre şu parametreler kullanılmıştır, oluşturulan üç kolonlu tablo, x ekseninde yer alması için tablodaki ilk komponent; y ekseni için tablodaki ikinci komponent sütunları; renk paleti olarak 13 rengin seçildiği değişken atanmıştır. Sonuç olarak elde edilen grafikler kaydedilmiştir.

**35442 Uzunluğunda Vektörlere Sahip Matris Kullanımı**

Bu kısımda görselleştirilmeler için kullanılacak verilerin yapılandırılmasında "**35542 Uzunluklu Vektörler İçeren Matrisin Oluşturulması"** bölümünde uygulanan akış uygulanmıştır.

Gen ifade, mutasyon, metilasyon, KSD veri tiplerinin görselleştirilmesi için ise yukarıda "**Kaynak Veri"** bölümünde izlenen aşamaların aynısı uygulanmıştır. Birleştirilmiş omik veri (35542 vektör uzunluğundaki) için de aynı şekilde önceden oluşturulan bir dosya olan GDSC_cell_line_reshape_transposed_minimized_16468_for_rforest_35K_v1.txt içindeki tablo değerlendirilerek t-SNE grafiği oluşturulmuştur.

**3747 Uzunluğundaki Vektörlere Sahip Matris**

GDSC_extracted_988_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform_v2.txt isimli önceden oluşturulmuş dosya veri görselleştirilmesinde kullanılmıştır. Gen ifade, mutasyon, metilasyon, KSD veri tipleri bu dosyadaki tablo üzerinden elde edilmiştir. Omik verilerin görselleştirilmesi için yukarıda "**Kaynak Veri"** bölümünde izlenen aşamaların aynısı uygulanarak t-SNE grafikleri oluşturulup kaydedilmiştir.

### 8.2.5. Veri İçi Alan Analizi

Bu uygulama kapsamında 3 farklı analiz tipi uygulanmıştır, hücre hattı, ilaç, rastgele bölümlendirme. Tüm analiz tiplerinde değerlendirilecek olan veri yapıları "1989 Uzunluklu Vektörler İçeren Matrisin Oluşturulması" bölümünde doku temelli olarak oluşturulan tablolardır. Belirlenen 13 ayrı doku için her üç analiz de yapılmıştır. Analiz tiplerinin kendi içlerinde tüm doku veri yapıları üzerinde aynı işlem aşamaları uygulandığı için aşağıdaki anlatımlar yalnızca birer doku (akciğer dokusu) üzerinden yapılmıştır.

**Hücre Hattı Özdeşliği Temelli Bölümlendirme (HHÖTB)**

Pandas, NumPy, Math kütüphaneleri hazır hale getirildikten sonra, hücre hattı ismi, ilaç parmak izi ve pIC50 değeri sütununa sahip olan GDSC_drug_response_df_cell_line_fingerprint_pIC50s_for_rforest.txt ve aynı tablonun parmak izi yerine ilaç ismini içeren versiyonu olan GDSC_drug_response_drugs_with_smiles_cell_lines_pIC50_v1.txt dosyasındaki tablolar birer değişkene atanmıştır. İlaç isimlerini içeren tablo hücre hattı isimleri üzerinden sıralanmış ve indeks sütunu yeniden 0 değerinden başlatılmıştır. Tablodaki ilaç isimleri bir listeye aktarılmıştır. Sonrasında, isim listesi yeni bir sütun olarak parmak izi içeren tabloya aktarılmıştır. Böylece, ilaç isimleri ve ilaç parmak izleriyle bir sözlük oluşturulabilmiştir.

Sklearn kütüphanesi içindeki şu modüller ve parantez içindeki metotları hazır hale getirilmiştir, preprocessing (StandardScaler), model_selection (train_test_split, GridSearchCV, cross_val_score), metrics, ensemble (RandomForestRegressor).

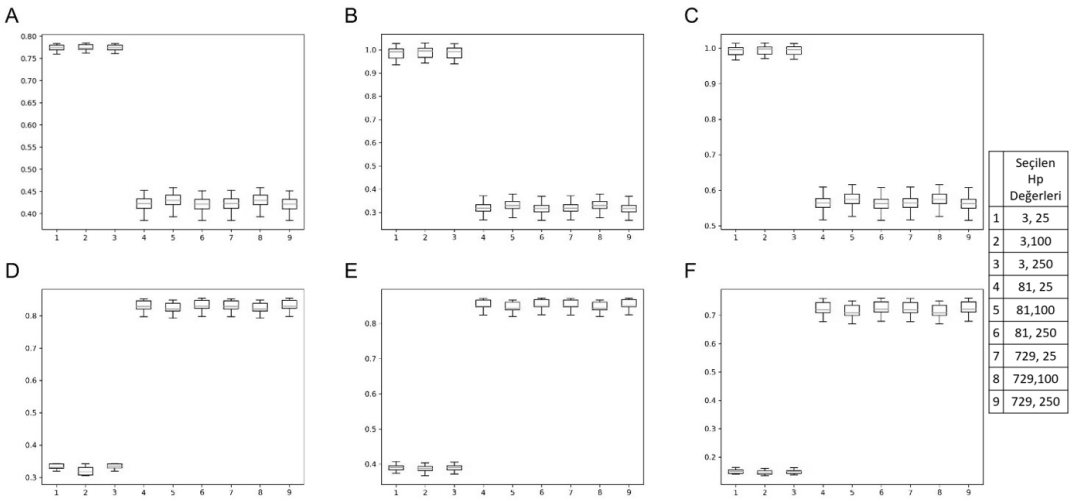
aero_digestive_tract_tissue_specific_cell_lines_L1000_commons_923_110_12_944_df.txt dosyasında bulunan akciğer dokusuna ait veri tablosu bir değişkene aktarılıp hücre hattına ait sütundaki değerler bir listede tutulmuştur.

Hücre hattı listesi, dört sütunlu ilaç yanıtı tablosundan ilgili hücre hatlarının filtrelenmesi için kullanılmıştır. Ardından, tabloda kalan ilaçların parmak izi değerleri bir listeye aktarılmıştır. Liste içindeki parmak izi değerleri bir döngü yardımıyla birbirinden ayrılıp aynı sırayla yeni bir liste içinde depolanmıştır. Random modülü hazır hale getirilip hücre hattı listesi rastgele şekilde sıralanmış ve sonradan aynı sıra elde edilmesi için 2 değeri verilmiştir. Ayrılan ilaç parmak izlerinin bulunduğu liste bir tablo haline getirilmiştir. Bu tabloya hücre hattı, ilaç isimleri ve pIC50 değerleri yeni sütunlar olarak eklenmiştir. Tablodaki sütun değerlerin her birine akciğer vektör matrisinin sütun sayısı değeri eklenmiştir.

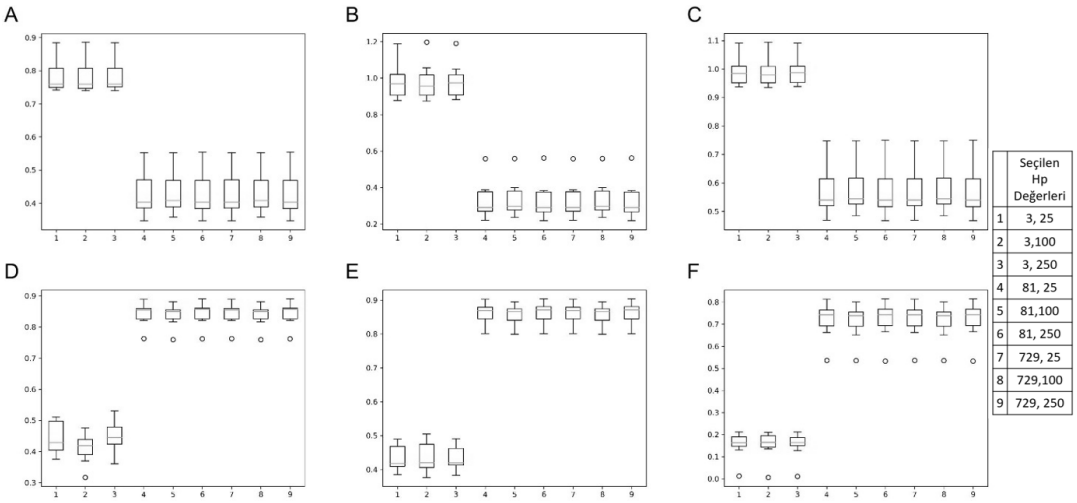
Akciğer vektör tablosu ve ayrılmış parmak izi tablosu hücre hattı isim sütunu üzerinden birleştirilmiştir. Oluşan tablodan, ilaç ve hücre hattı isim sütunları çıkarılıp birer listeye alınmıştır. Os modülü, Scipy kütüphanesinin stats modülünden spearmanr, pearsonr metotları hazır hale getirilmiştir. Bir boş dosya ve skorlama parametreleri için sütun isimleri oluşturulup kaydedilmiştir. Model hiperparametrelerinde kullanılacak 3'er değer şu şekilde seçilmiştir, max_split (3, 81, 729), n_estimators (25, 100, 250).

Birleştirilmiş verinin modelde kullanılacak değişkenleri bulunduran sütunlar içinden pIC50; sonuç değerleri belirten veri için ise ilaç, hücre hattı isimleri ve pIC50 sütunları kullanılmıştır. Bir döngü içinde hücre hattı listesinin yüzde 10'u test verisi için; kalan yüzde 90'I ise eğitim verisi için yeni listelere aktarılmıştır. Değişkenler verisi ve sonuçlar verisi üzerinde bu iki liste ile filtrelemeler yapılıp eğitim ve test setleri oluşturulmuştur. Sonrasında, ilaç ve hücre hattı isim sütunları bu setlerden çıkartılmıştır.

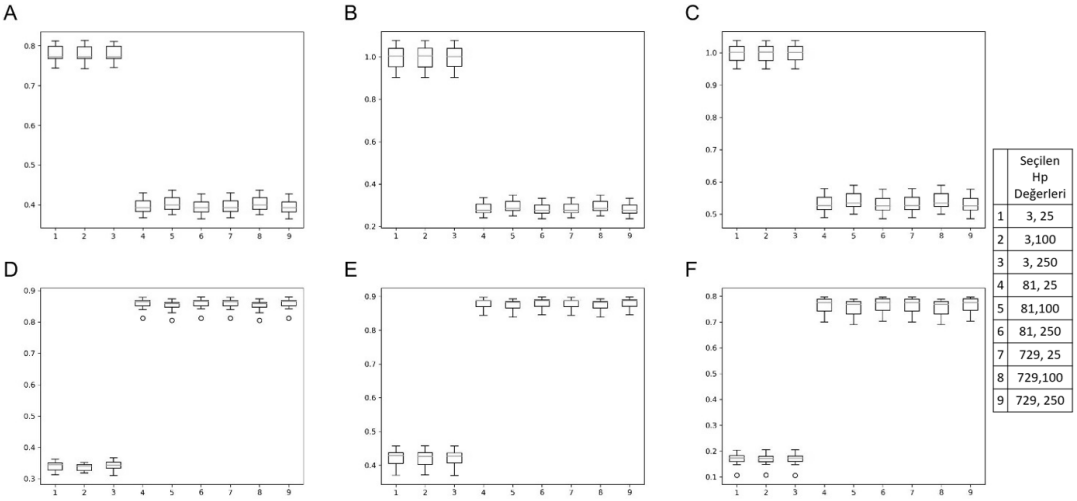
İki iç içe döngü, her ikili hiperparametre kombinasyonunun modellenebilmesi için oluşturulmuştur. Her döngü adımında, RandomForestRegressor metodu içinde max_depth, n_estimators, verbose=3 (tüm çıktıların yazdırılması için), n_jobs=6 (kullanılan iş paketi sayısı), random_state=2) parametreleri kullanılarak bir obje oluşturulmuş ve eğitim verisi bu obje yardımıyla modele uydurulmuştur. Model içinde değerlendirilmesi için test verisi için tahmin değerleri oluşturulmuştur. Her döngü adımında oluşturulan boş dosyaya tahmin ve gerçek değerler arasındaki fark MAE, MSE, RMSE, SCC, PCC, R^2 skorlama metrikleri ile hesaplanıp yazdırılmıştır. Her döngüde tahmin edilen değerler ve gerçek değerler hücre hattı ve ilaç çifti isimleriyle beraber yazdırılıp döngü bitiminde bir dosyaya yazdırılıp kaydedilmiştir. Şekil 8.6 – 8.18'de, HHÖTB tahmin performanslarının görselleştirilmesi için dokulara özel olarak kutu grafikleri hazırlanmıştır. Bu şekillerde, her skorlama metriğine ait birer alt şekil bulunmaktadır. Alt şekillerdeki yatay eksenlerde, seçilen hiperparametrelere ait değerlerinin ikili kombinasyonları (9 adet) belirtilmiştir.



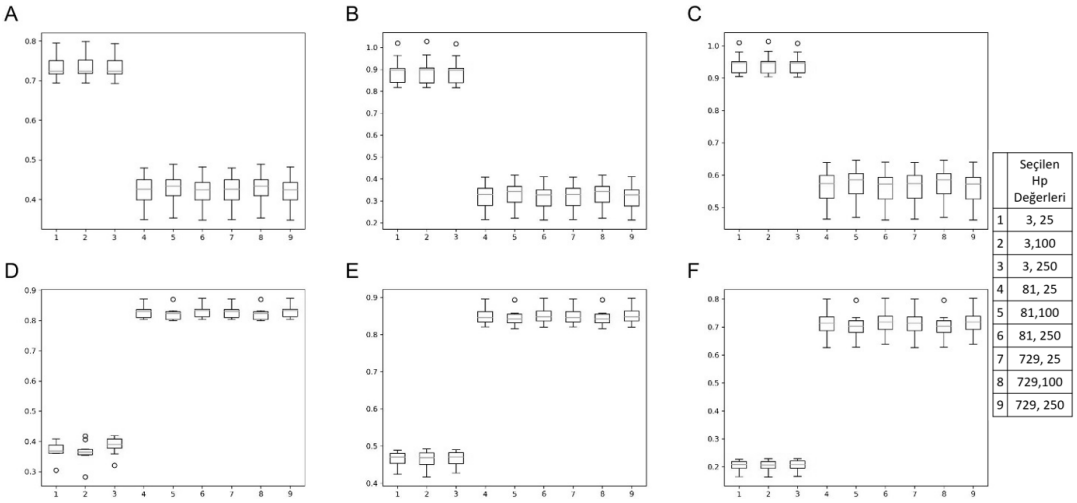

**Şekil 8.6.** Akciğer dokusu hücre hatlarıyla yapılan HHÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



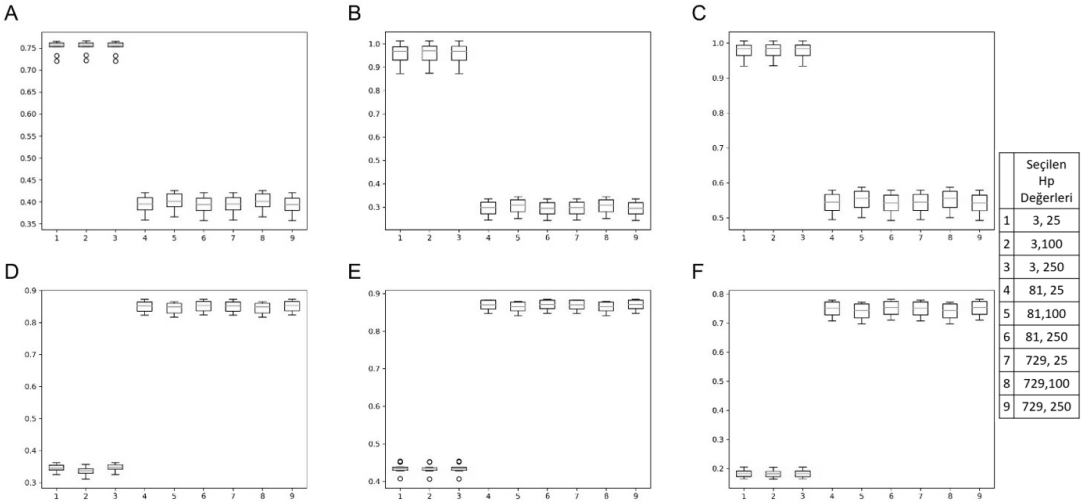

**Şekil 8.7.** Böbrek dokusu hücre hatlarıyla yapılan HHÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



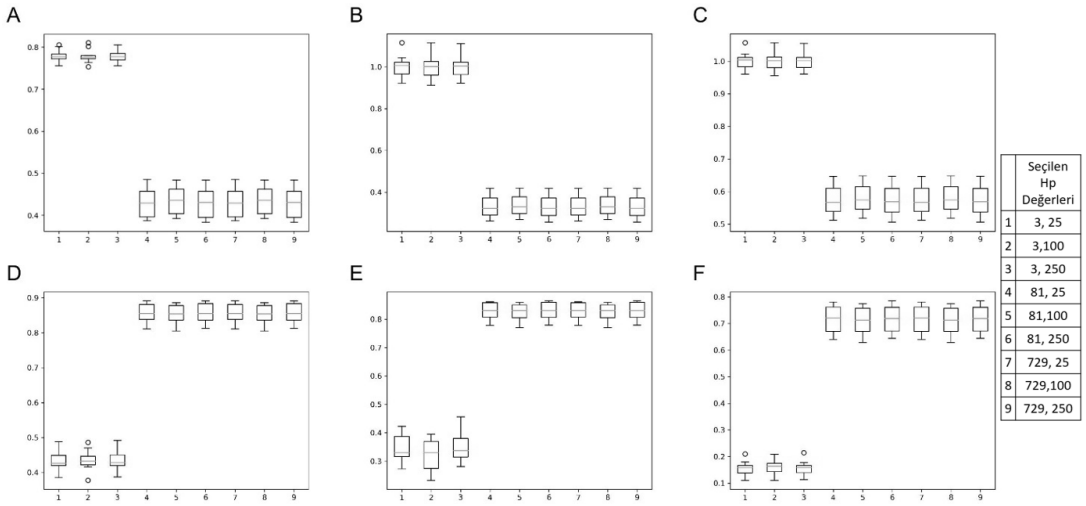

**Şekil 8.8.** Deri dokusu hücre hatlarıyla yapılan HHÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



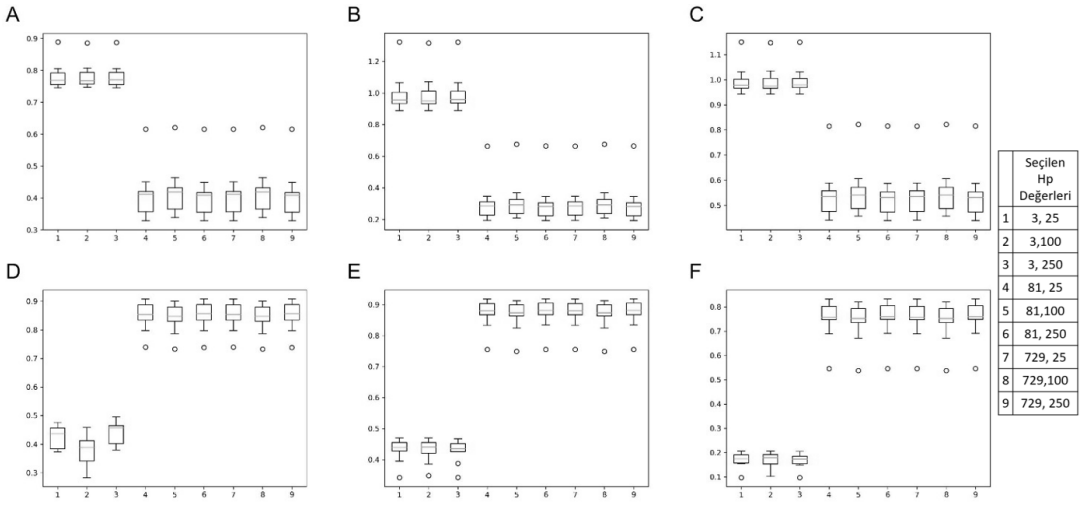

**Şekil 8.9.** Meme dokusu hücre hatlarıyla yapılan HHÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



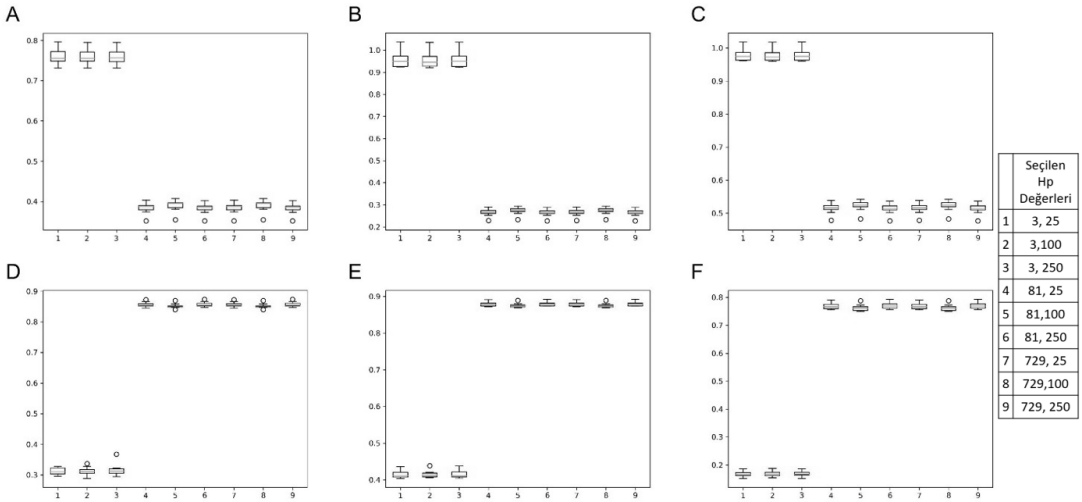

**Şekil 8.10.** Kan dokusu hücre hatlarıyla yapılan HHÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



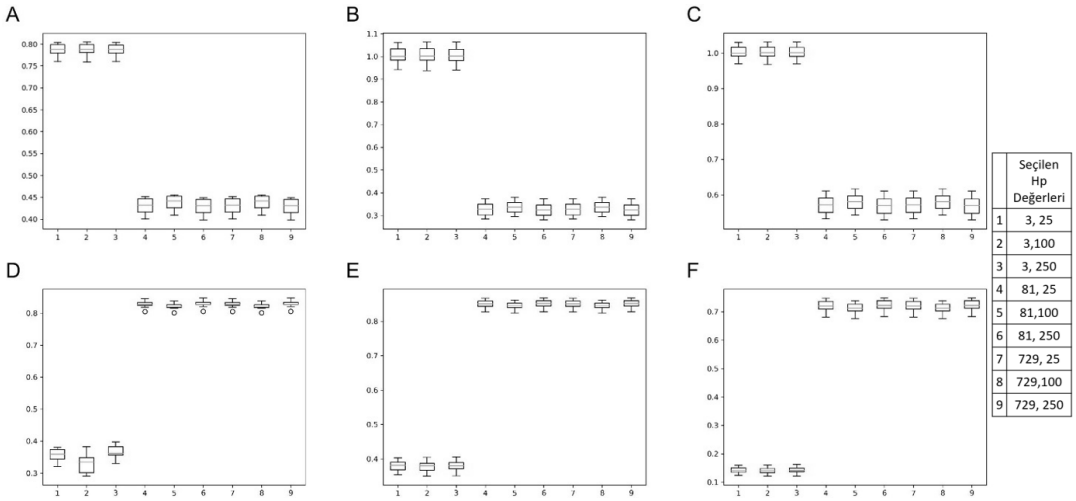

**Şekil 8.11.** Kemik dokusu hücre hatlarıyla yapılan HHÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



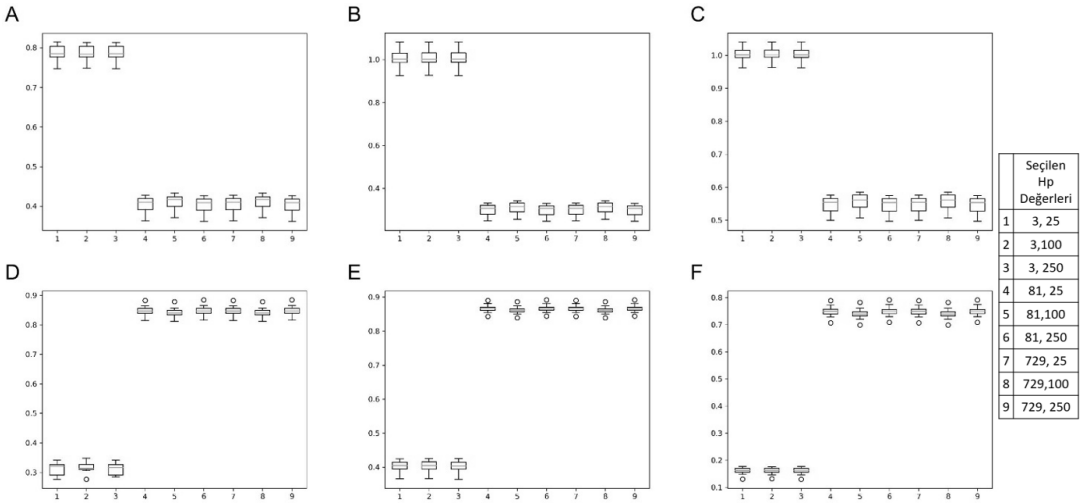

**Şekil 8.12.** Pankreas dokusu hücre hatlarıyla yapılan HHÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



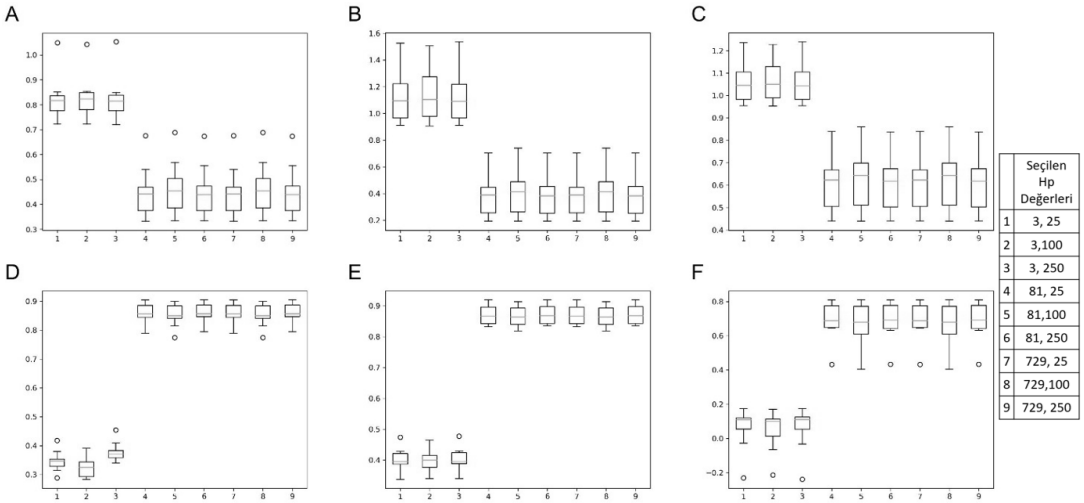

**Şekil 8.13.** Sinir sistemi dokusu hücre hatlarıyla yapılan HHÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



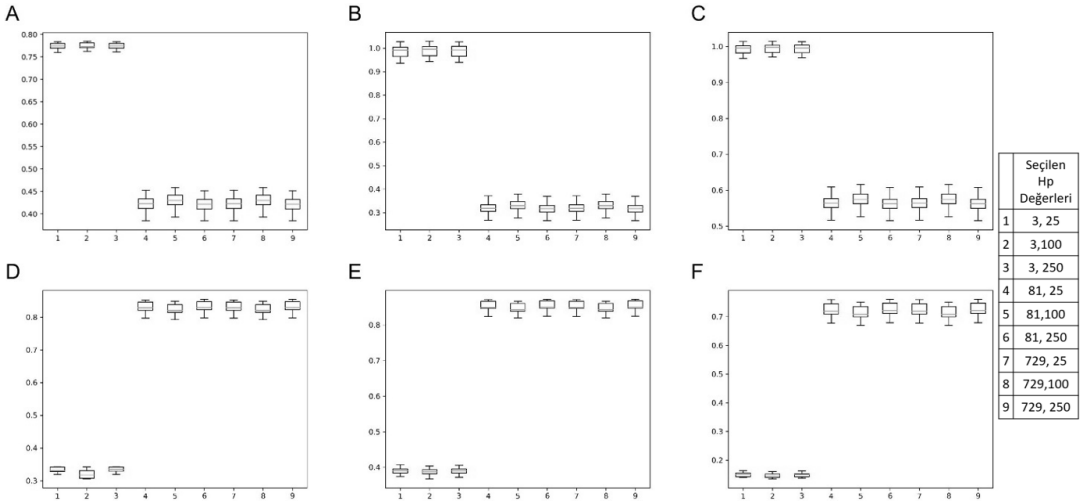

**Şekil 8.14.** Sindirim sistemi dokusu hücre hatlarıyla yapılan HHÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



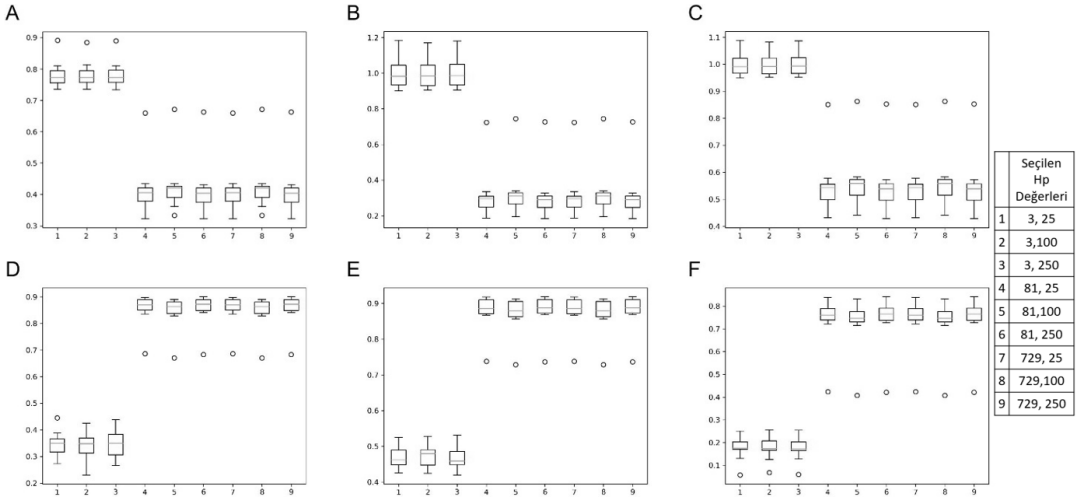

**Şekil 8.15.** Solunum-sindirim sistemi dokusu hücre hatlarıyla yapılan HHÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.




**Şekil 8.16.** Tiroid dokusu hücre hatlarıyla yapılan HHÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.




**Şekil 8.17.** Ürogenital sistem dokusu hücre hatlarıyla yapılan HHÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



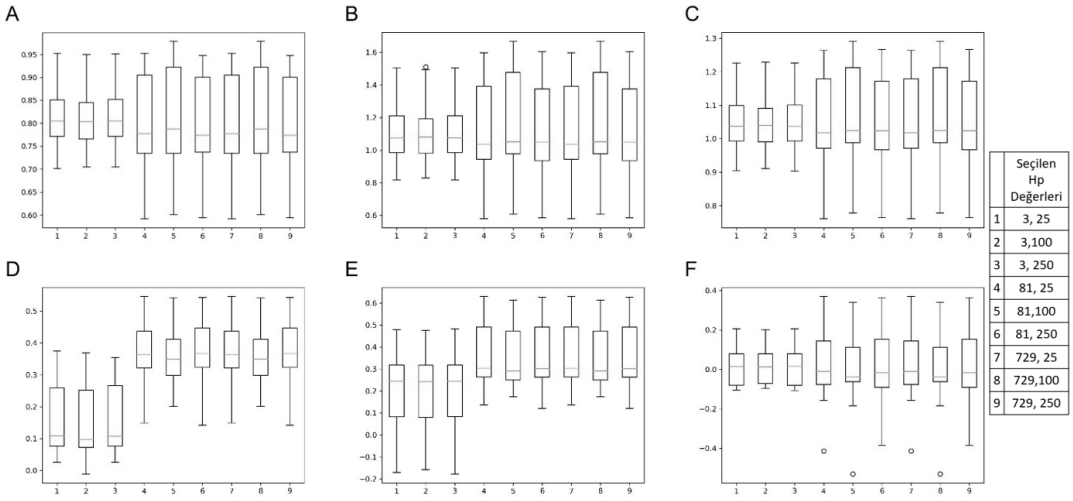

**Şekil 8.18.** Yumuşak doku hücre hatlarıyla yapılan HHÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.

### İlaç Özdeşliği Temelli Bölümlendirme (İÖTB)

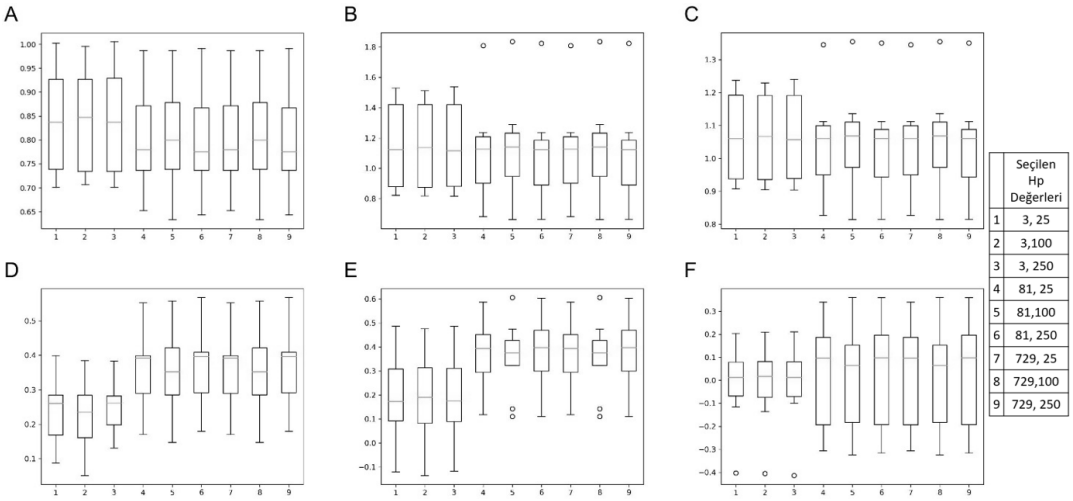
İÖTB'de de HHÖTB için uygulanan prosedürlerin hemen hemen bir benzeri uygulanmıştır. Farklılıklar aşağıda belirtilmiştir.

HHÖTB'de uygulanan hücre hattı listesinin rastgele şekilde sıralanması aşaması yerine, ilaç listesinin sıralanması yapılmıştır. Bunun devamında da, eğitim ve test verisinin oluşturulması aşamasında yeni oluşturulan bu ilaç listesi kullanılmıştır.

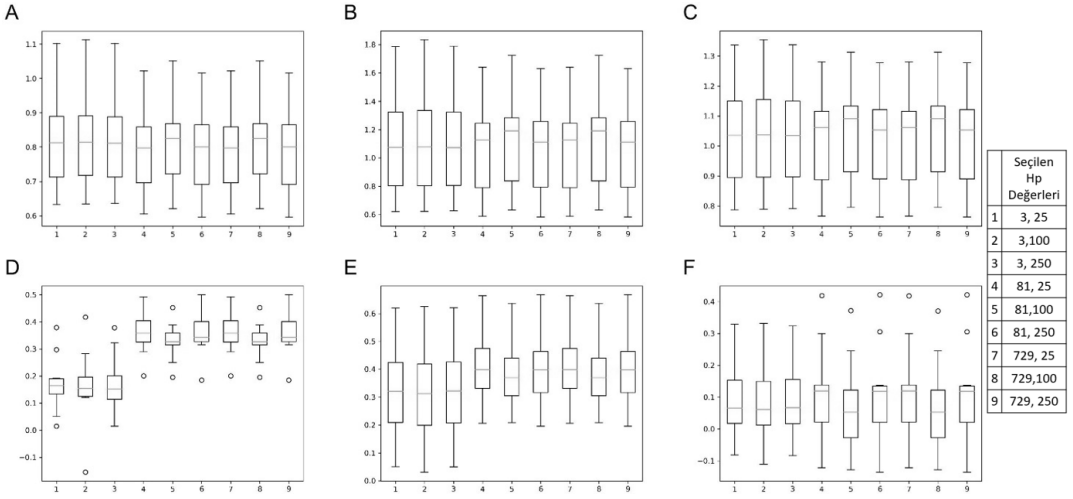
Şekil 8.19 – 8.31'de, İÖTB tahmin performanslarının görselleştirilmesi için dokulara özel olarak kutu grafikleri hazırlanmıştır. Bu şekillerde, her skorlama metriğine ait birer alt şekil bulunmaktadır. Alt şekillerdeki yatay eksenlerde, seçilen hiperparametrelere ait değerlerinin ikili kombinasyonları (9 adet) belirtilmiştir.



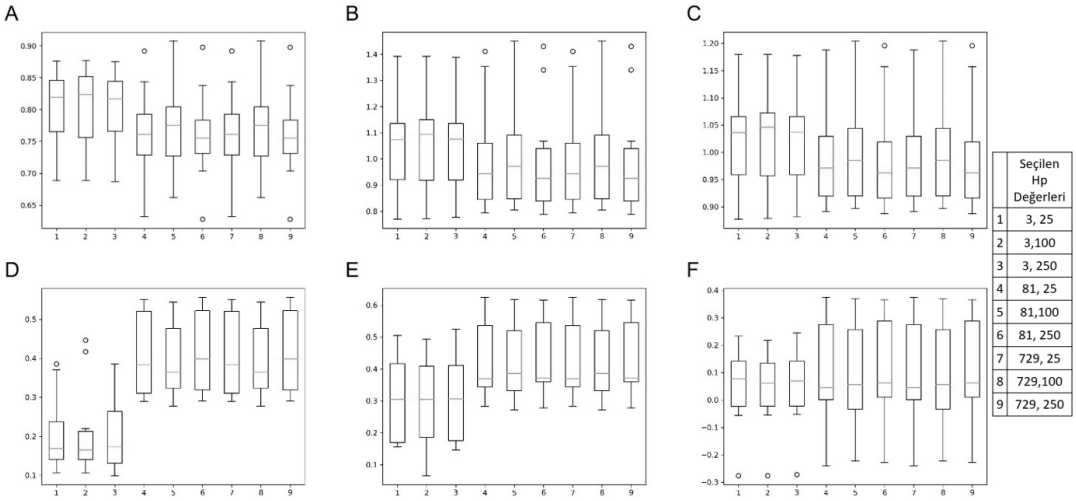

**Şekil 8.19.** Akciğer dokusu hücre hatlarıyla yapılan İÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



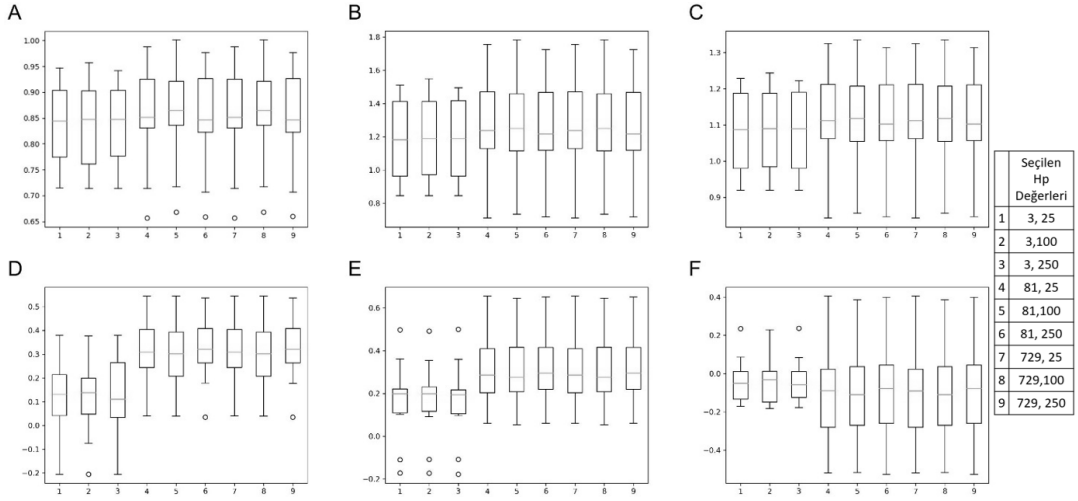

**Şekil 8.20.** Böbrek dokusu hücre hatlarıyla yapılan İÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



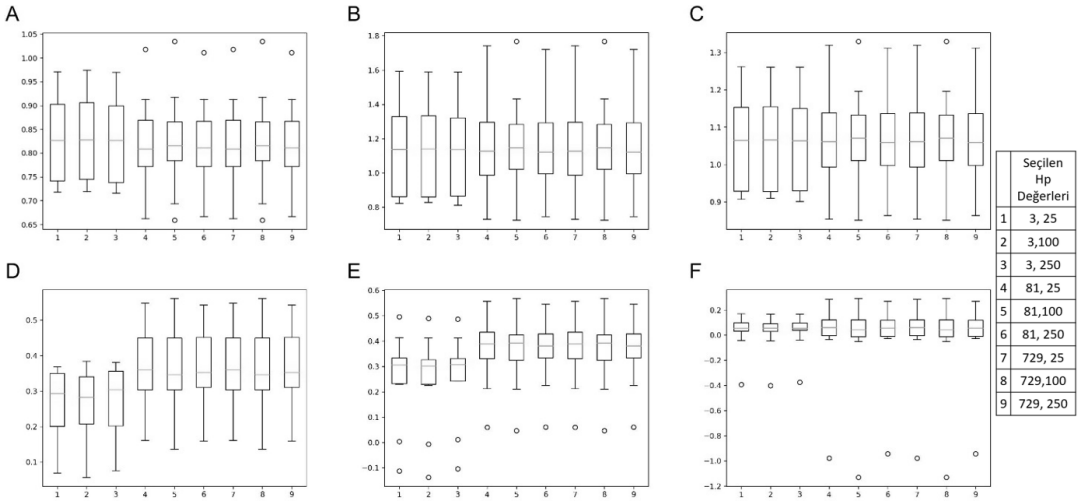

**Şekil 8.21.** Deri dokusu hücre hatlarıyla yapılan İÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



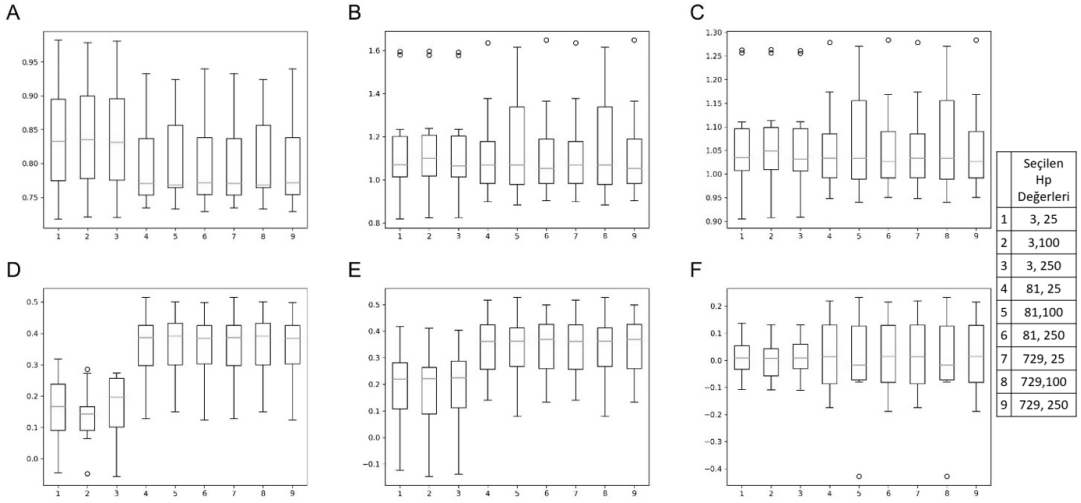

**Şekil 8.22.** Meme dokusu hücre hatlarıyla yapılan İÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



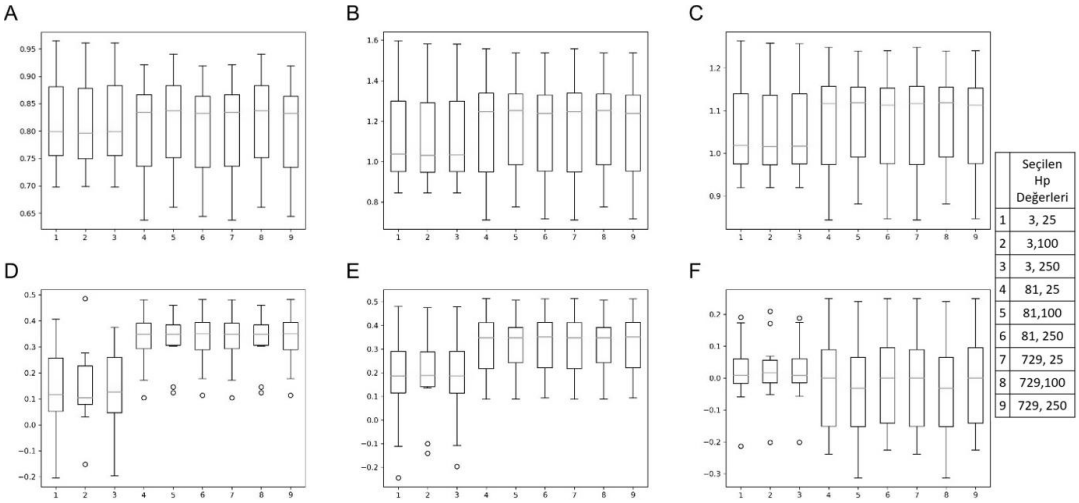

**Şekil 8.23.** Kan dokusu hücre hatlarıyla yapılan İÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



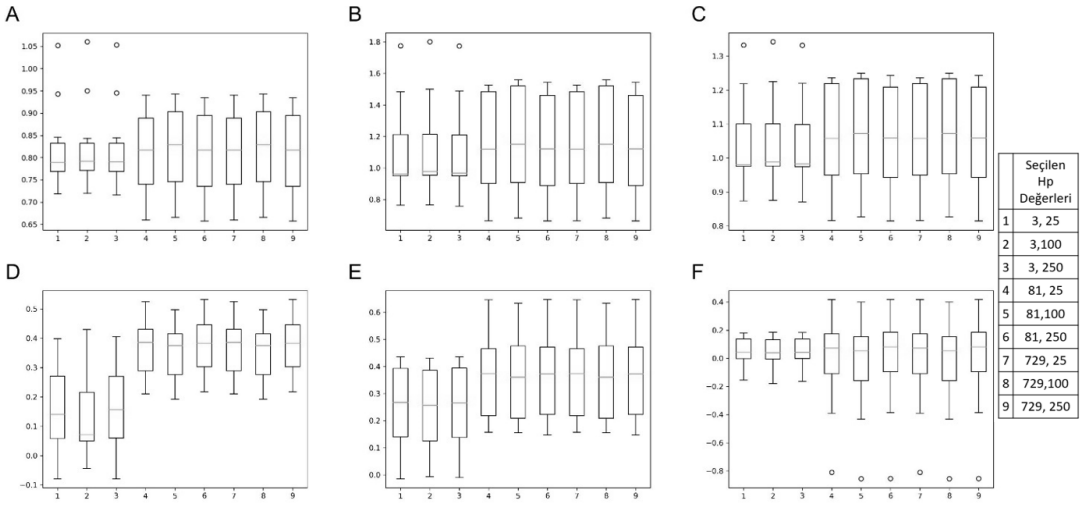

**Şekil 8.24.** Kemik dokusu hücre hatlarıyla yapılan İÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



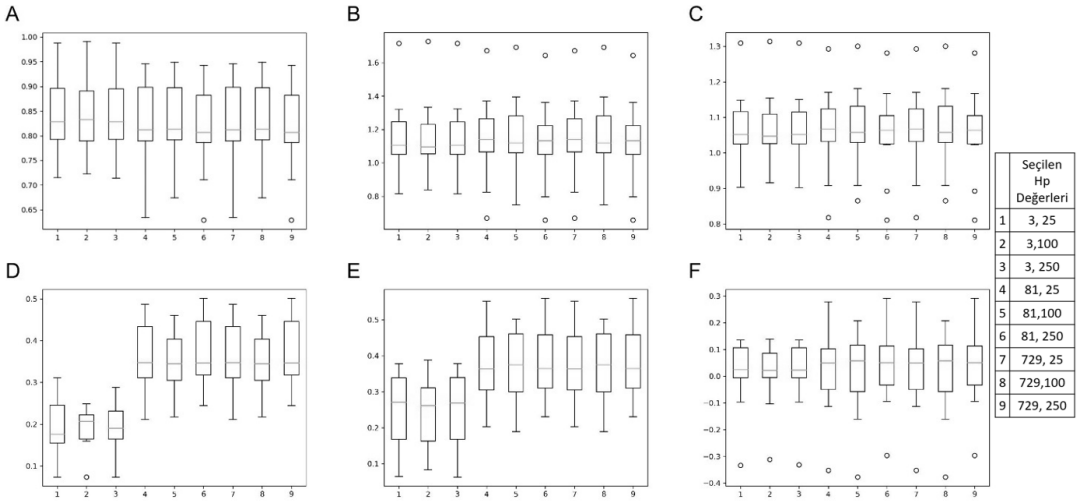

**Şekil 8.25.** Pankreas dokusu hücre hatlarıyla yapılan İÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



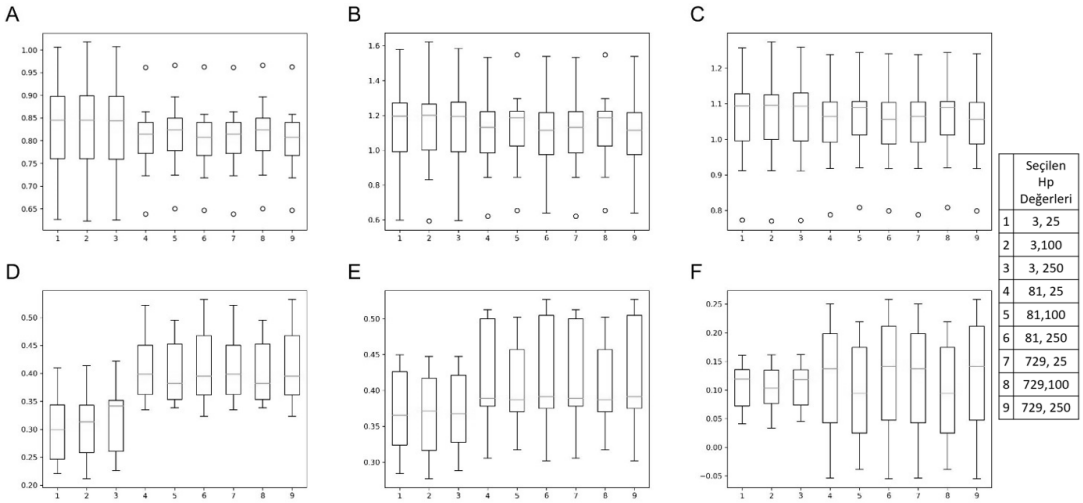

**Şekil 8.26.** Sinir sistemi dokusu hücre hatlarıyla yapılan İÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



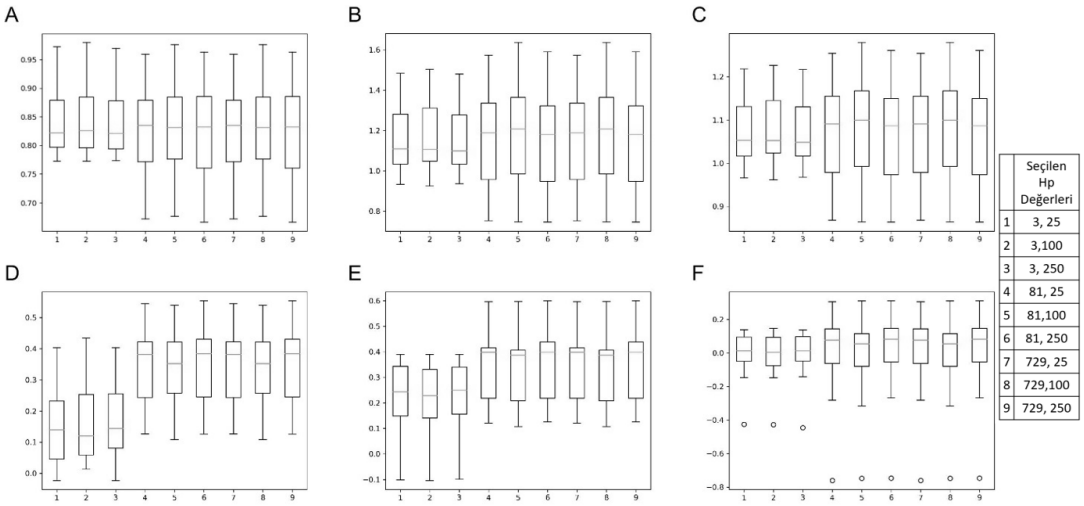

**Şekil 8.27.** Sindirim sistemi dokusu hücre hatlarıyla yapılan İÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



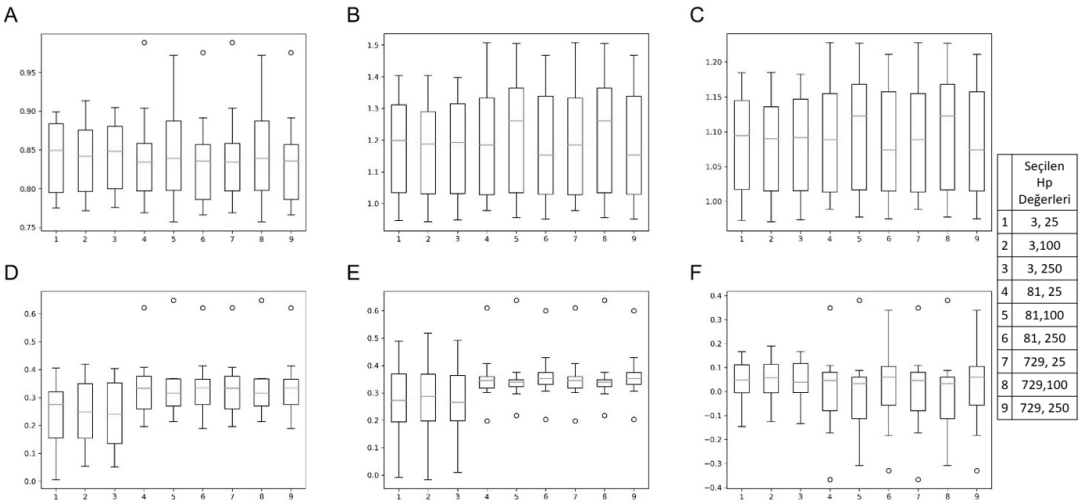

**Şekil 8.28.** Solunum-sindirim sistemi dokusu hücre hatlarıyla yapılan İÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.




**Şekil 8.29.** Tiroid dokusu hücre hatlarıyla yapılan İÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.




**Şekil 8.30.** Ürogenital sistem dokusu hücre hatlarıyla yapılan İÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.




**Şekil 8.31.** Yumuşak doku hücre hatlarıyla yapılan İÖTB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.

**Rastgele Bölümlendirme (RB)**

RB'de de HHÖTB için uygulanan prosedürlerin bir benzeri uygulanmıştır. Farklılıklar aşağıda belirtilmiştir. Veri, özdeşlik üzerinden bölümlendirilmeyeceği için HHÖTB'de uygulanan hücre hattı listesinin rastgele şekilde sıralanması aşaması çıkartılmıştır.

Sklearn içinde yer alan metrics modülündeki make_scorer ve SCORERS metotları hazır hale getirilmiştir. SCC ve PCC metriklerinin çapraz doğrulama için Sklearn skorlama metriklerinin arasında olmaması nedeniyle birer fonksiyon oluşturulup make_scorers metodu yardımıyla SCORERS içine dahil edilmiştir.

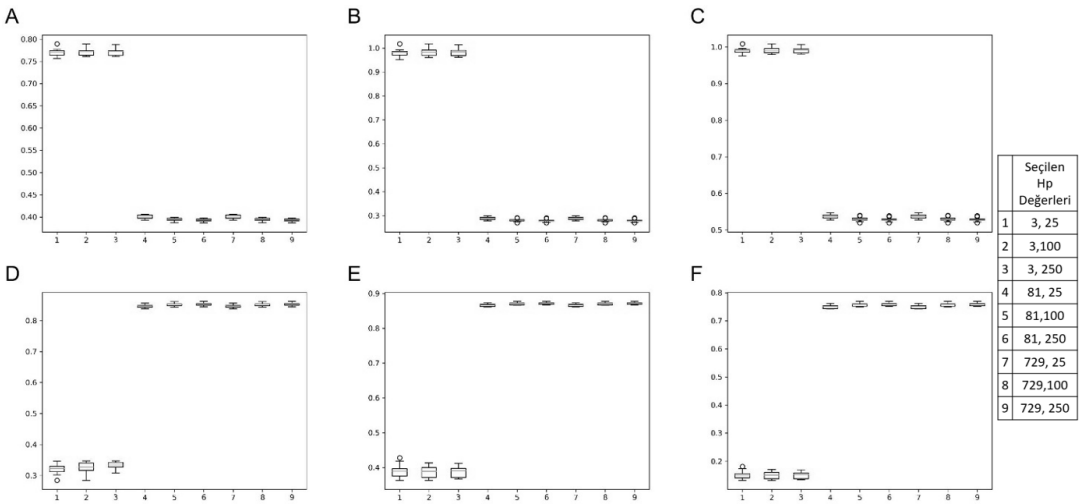
Sklearn'in model_selection modülündeki cross_validate, cross_val_predict, KFold metotları hazır hale getirilmiştir. Diğer iki bölümlendirme yönteminden farklı olarak, cross_val_predict ile 10 kat çapraz doğrulama uygulamasından tahmin sonuçları alınmıştır. Ek olarak, cross_validate metodu ile 10 kat çapraz doğrulama yapılıp skorlama metrikleriyle tahmin performansı ölçülmüştür.

cross_val_score sonuçlarının kaydedilmesi için boş bir dosya oluşturulup dosya içine hücre hattı – ilaç çiftinin deneysel ilaç yanıtı ve tahmin sonucu değeri başlıkları eklenmiştir. KFold metodu ile çapraz doğrulama için 10 kat uygulama ve rastgelelik için 2 sayısı verilmiş ve bir obje üzerine kaydedilmiştir. HHÖTB'de olduğu gibi max_depth ve n_estimators hiperparametreleri için aynı değerler kullanılmıştır. İç içe iki döngü yapılarak bu değerlerin ikili kombinasyonları için iki farklı çapraz doğrulama işlemi yapılmıştır.

RandomForestRegressor metoduyla HHÖTB'deki uygulamadan farklı olarak n_jobs için 4 değeri kullanılarak iş paketi sayısı azaltılmıştır. Birinci çapraz doğrulama işlemi, cross_val_predict ile yapılıp test setleri için tahmini değerler yaratılmıştır. Her döngü adımında elde edilen sonuçlar hücre hattı – ilaç çifti isimleri ve gerçek GDSC değeriyle beraber önceden oluşturulan boş dosya içine kaydedilmiştir. İkinci çapraz

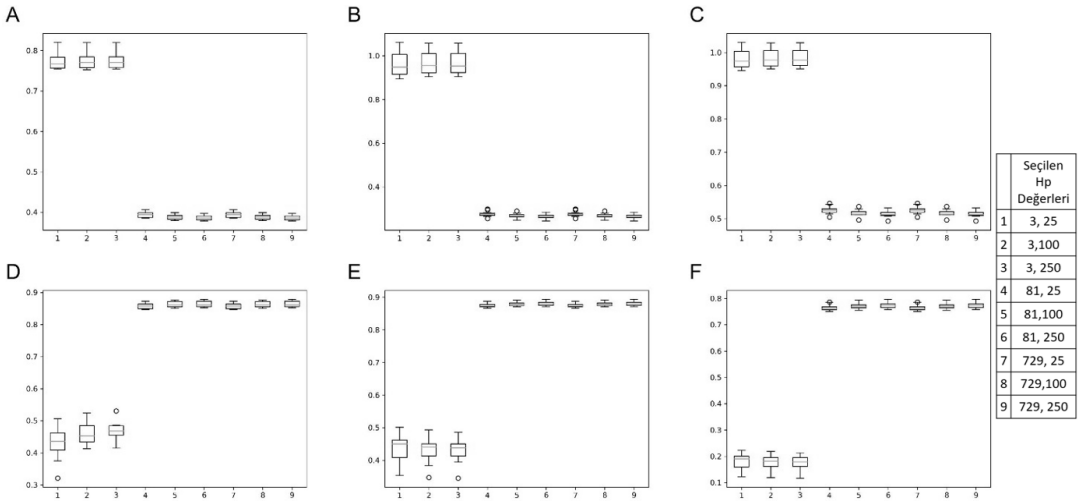
doğrulama yönteminde ise, her doğrulama adımında elde edilen sonuçlar bir dosyaya; bunların ortalamaları ise başka bir dosyaya metrik isimleri ile beraber kaydedilmiştir.

Şekil 8.32 – 8.44'de, RB tahmin performanslarının görselleştirilmesi için dokulara özel olarak kutu grafikleri hazırlanmıştır. Bu şekillerde, her skorlama metriğine ait birer alt şekil bulunmaktadır. Alt şekillerdeki yatay eksenlerde, seçilen hiperparametrelere ait değerlerinin ikili kombinasyonları (9 adet) belirtilmiştir.

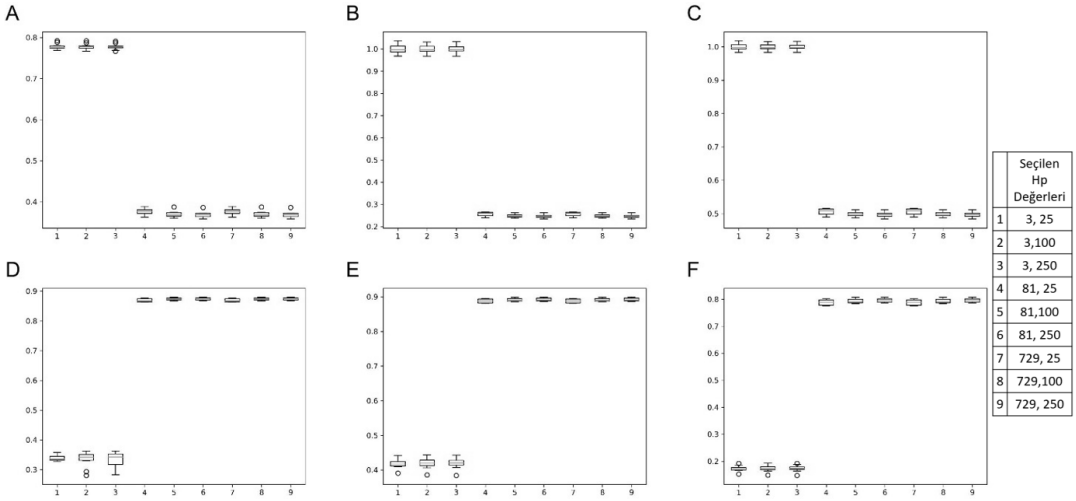



| | Seçilen Hp Değerleri |
|---|---|
| 1 | 3, 25 |
| 2 | 3,100 |
| 3 | 3, 250 |
| 4 | 81, 25 |
| 5 | 81,100 |
| 6 | 81, 250 |
| 7 | 729, 25 |
| 8 | 729,100 |
| 9 | 729, 250 |

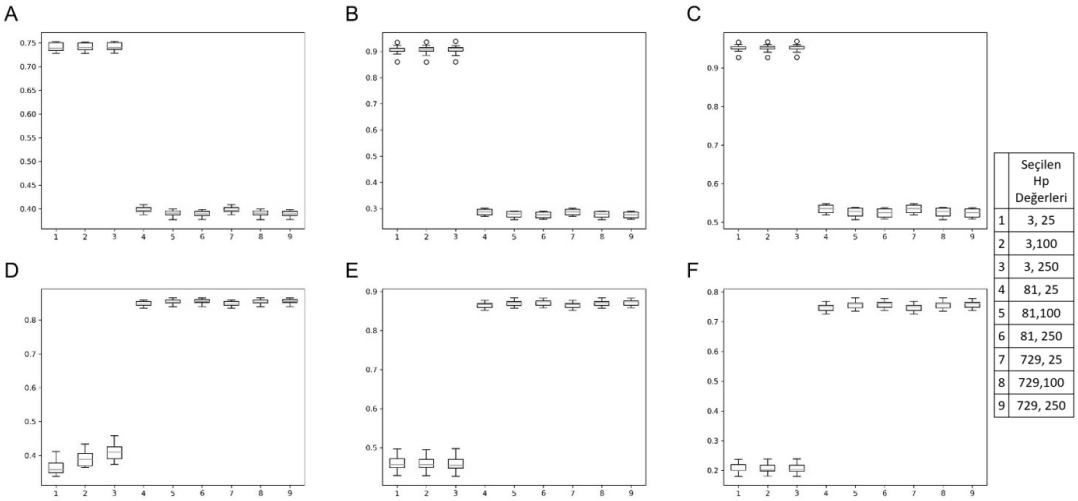
**Şekil 8.32.** Akciğer dokusu hücre hatlarıyla yapılan RB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



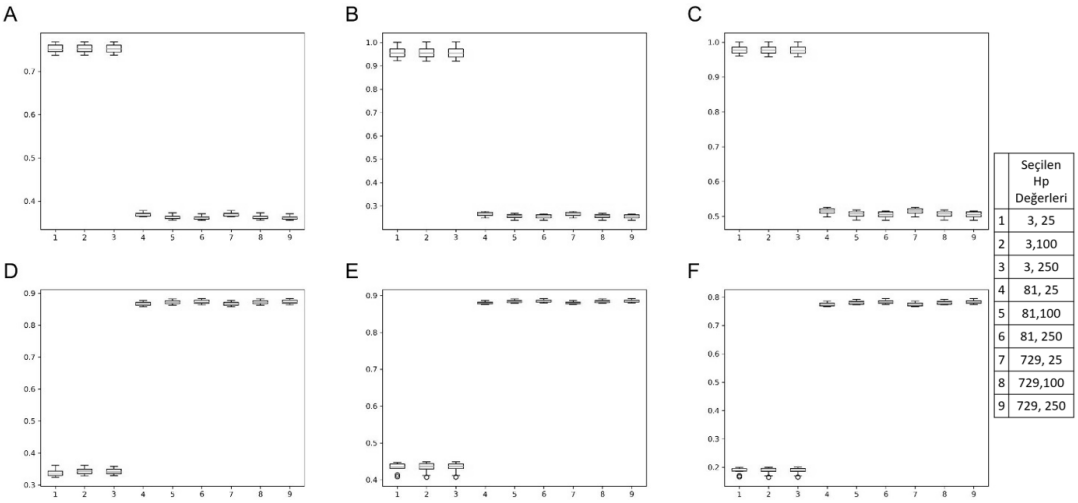

**Şekil 8.33.** Böbrek dokusu hücre hatlarıyla yapılan RB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



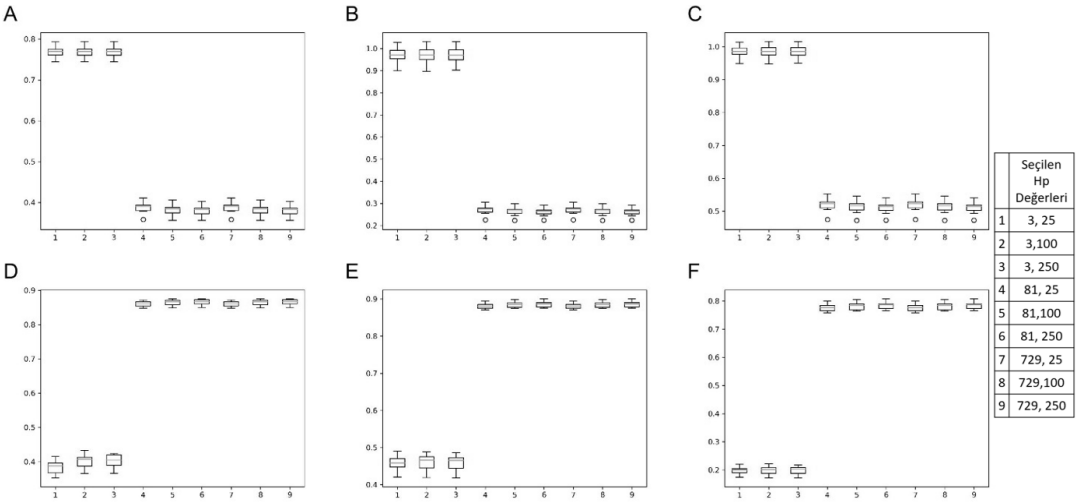

**Şekil 8.34.** Deri dokusu hücre hatlarıyla yapılan RB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



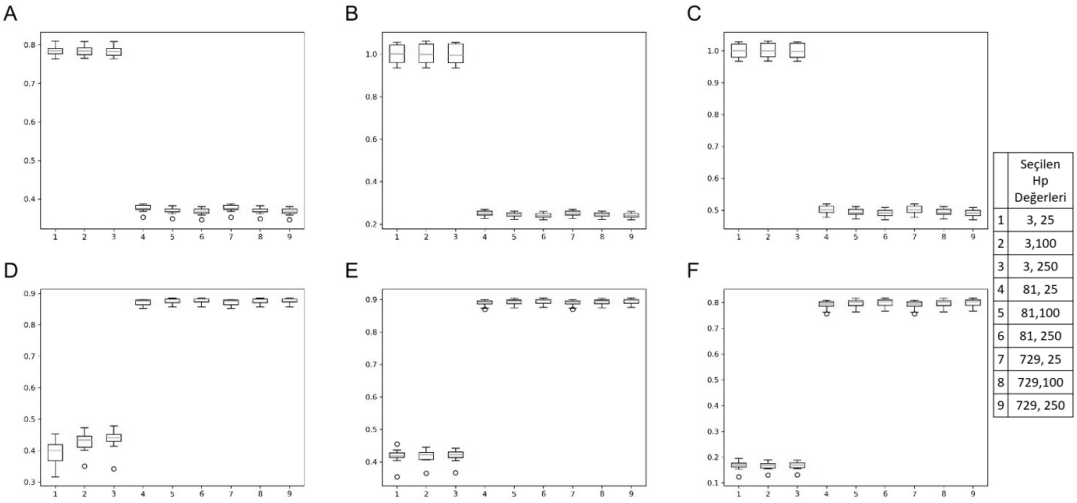

**Şekil 8.35.** Meme dokusu hücre hatlarıyla yapılan RB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



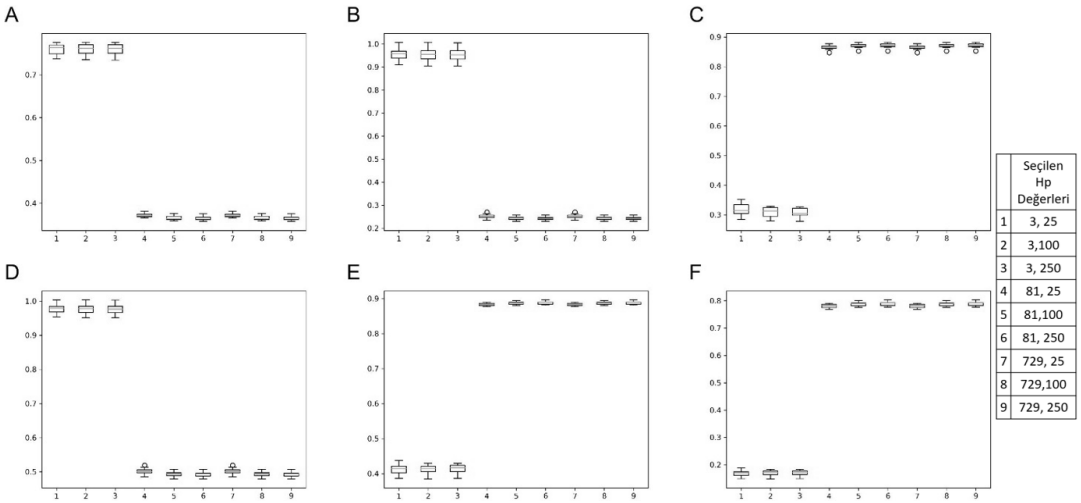

**Şekil 8.36.** Kan dokusu hücre hatlarıyla yapılan RB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



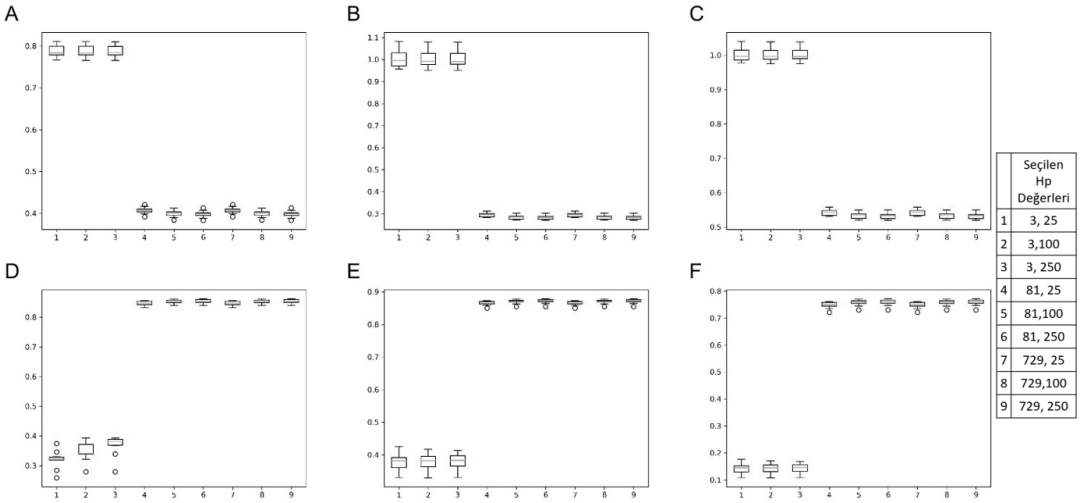

**Şekil 8.37.** Kemik dokusu hücre hatlarıyla yapılan RB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



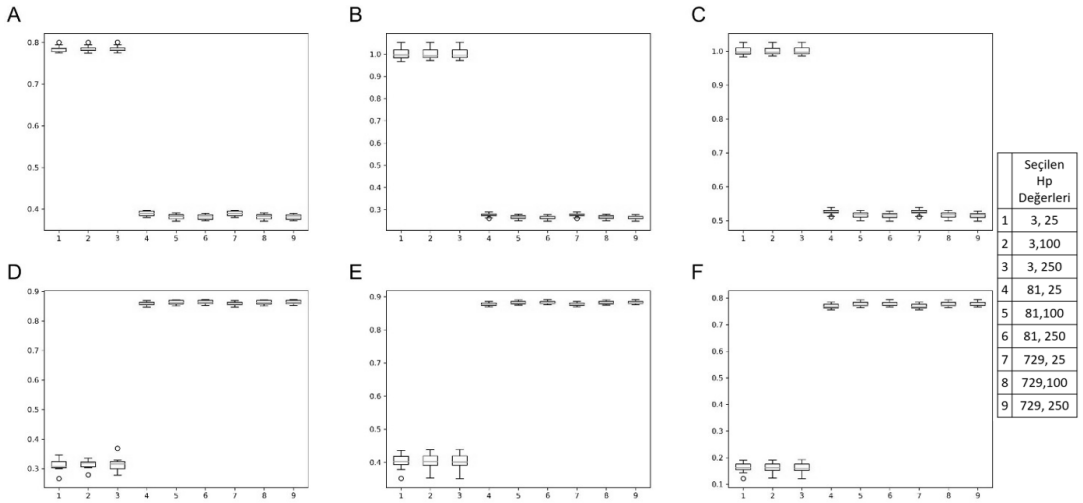

**Şekil 8.38.** Pankreas dokusu hücre hatlarıyla yapılan RB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



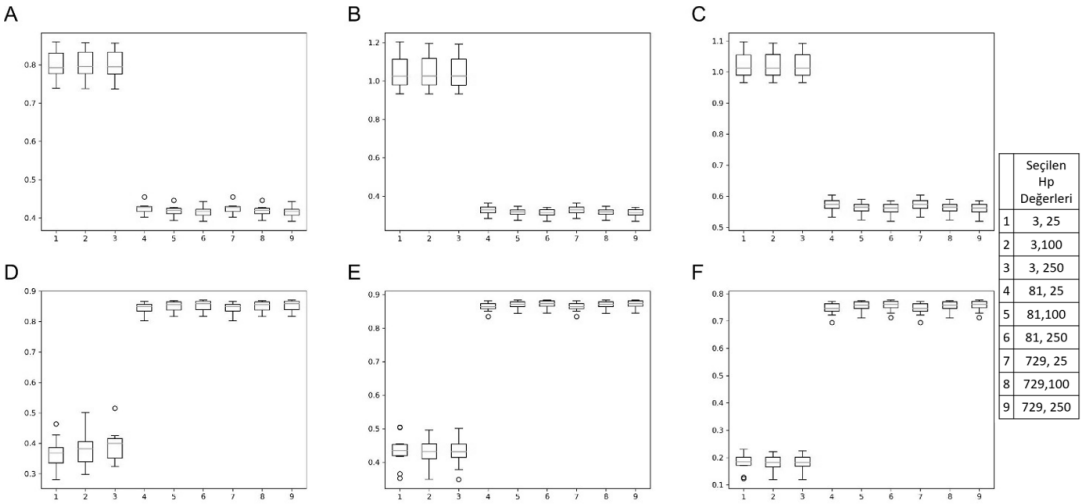

**Şekil 8.39.** Sinir sistemi dokusu hücre hatlarıyla yapılan RB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



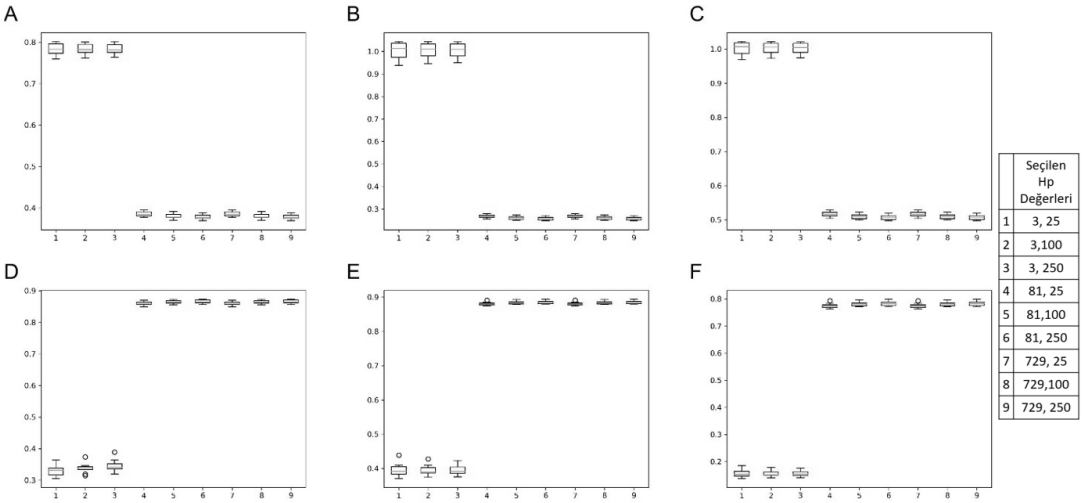

**Şekil 8.40.** Sindirim sistemi dokusu hücre hatlarıyla yapılan RB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.



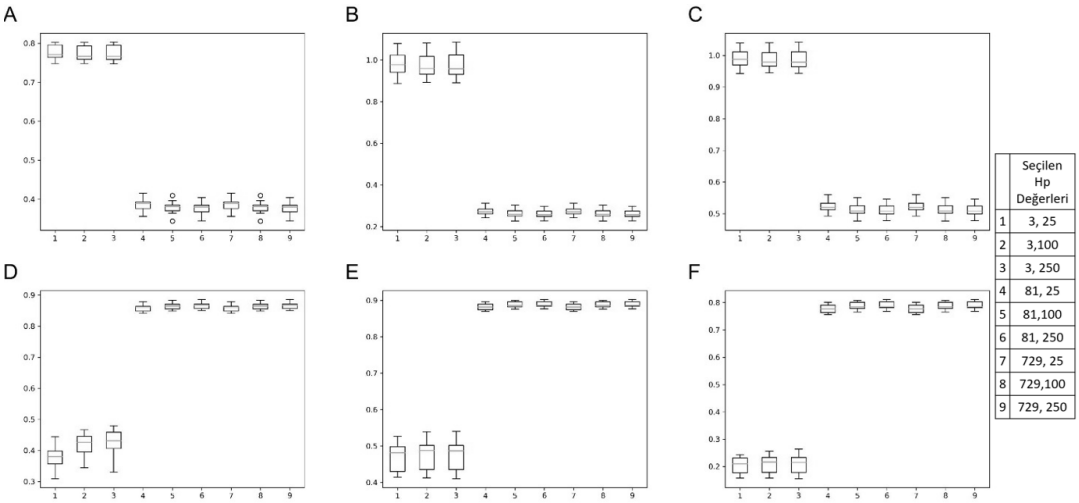

**Şekil 8.41.** Solunum-sindirim sistemi dokusu hücre hatlarıyla yapılan RB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.




**Şekil 8.42.** Tiroid dokusu hücre hatlarıyla yapılan RB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.




**Şekil 8.43.** Ürogenital sistem dokusu hücre hatlarıyla yapılan RB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.




**Şekil 8.44.** Yumuşak doku hücre hatlarıyla yapılan RB analizi sonuçlarına ait kutu grafikleri. Alt şekillerdeki yatay eksenlerde yer alan 1-9 arası değerler, max_depth (3, 81, 729) ve n_estimators (25, 100, 250) hiperparametrelerine atanan değerlerin ikili kombinasyonlarına denk gelmektedir. Tüm kombinasyonlar için elde edilen değerler, (A) MAE, (B) MSE, (C) RMSE, (D) SCC, (E) PCC, (F) $R^2$ skorlama metriklerine ait sonuçlar bir arada olacak şekilde verilmiştir.

## 8.3. EK-3

### 8.3.1. Çapraz Alan Analizi İçin GDSC Veri Tipleri Üzerinde Yapılan Ön İşlemler

#### Gen, Hücre Hattı ve İlaç İsimlerinin Düzenlenmesi

GDSC'deki 4 özellik veri tipi ve ilaç yanıtı verisi göz önüne alınarak gen ve hücre hatları isimleri üzerinde harf ve sayı dışındaki karakterlerin çıkarılıp küçük harfler ile yeniden düzenlenmesi işlemi, diğer platformlardaki isimlerle uyumlu olması için gerçekleştirilmiştir.

Düzenlenecek olan isim listeleri önceki adımlarda oluşturulan düzenlenmiş kaynak veri yapıları üzerinden alınmıştır.

Hücre hatları için, "NCI" kodunu içeren isimlerden bu kod çıkarılmıştır. "7860", "NTERASclD1", "KM-H2", "KMH-2", "TT", "T-T", "SC-1" isimleri sırasıyla "786O","ntera2cld1", "km-h2", "kmh-2", "tt" , "t-t", "sci1"ile değiştirilmiştir.

İlaç isimleri için, "PFI-3", "PFI3" isimleri sırasıyla "pfi-3", "pfi3 olarak değiştirilmiştir.

Düzenlenmiş halleri aynı olan isimler olduğunda bu orijinal isimlerin sadece küçük harfli ve çıkarılacak karakterleri korunmuş halleri kullanılmıştır.

Sonuç olarak, tüm veri yapıları kullanılarak en geniş haliyle 1155 hücre hattı ismi, 29870 gen ismi bulunmuştur ve tablo genişletme sırasında kullanılacak ana listeler böylece belirlenmiştir.

Tüm özellik tipleri ve ilaç yanıtı tablosu için oluşturulan sözlükler tablo halinde yazdırılıp kaydedilmiştir.

**Gen İfade Verisi**

Gen ifade verisi için önceki aşamalarda oluşturulan düzenlenmiş kaynak veri kullanılarak hücre hattı ve gen isimlerinin eklenmesiyle tabloda genişletme işlemi yapılmıştır.

Verideki orijinal isimler düzenlenmiş halleriyle değiştirilmiştir.

GDSC için dört özellik veri tipi baz alarak oluşturulan en geniş hücre hattı ve gen listelerinde olup gen ifade verisi üzerinde olmayan isimler ilişkili eksenlere eklenmiştir. Yeni oluşan tablo boşluklarına boşluk değeri (NaN, Not a Number) değeri atanmıştır. Toplamda 12451 gen ve 141 hücre hattı tabloya eklenmiştir.

Tablo eksenleri alfabetik sıralandıktan sonra sonuç tablosu olarak yazdırılmıştır. Sonuç tablosunda 29870 gen ve 1155 hücre hattı bulunmaktadır.

**Mutasyon Verisi**

Mutasyon verisi için önceki aşamalarda oluşturulan düzenlenmiş kaynak veri kullanılarak hücre hattı ve gen isimlerinin eklenmesiyle tabloda genişletme işlemi yapılmıştır.

Verideki orijinal isimler düzenlenmiş halleriyle değiştirilmiştir.

GDSC için dört özellik veri tipi baz alarak oluşturulan en geniş hücre hattı ve gen listelerinde olup gen ifade verisi üzerinde olmayan isimler ilişkili eksenlere eklenmiştir. Yeni oluşan tablo boşluklarına 0 (mutasyon yok) değeri atanmıştır. Toplamda 7901 gen ve 126 hücre hattı tabloya eklenmiştir.

Tabloda oluşan 29873 gen ismi olduğu görülünce birden çok olan gen isimlerinin ilk örnekleri tutulup diğerleri silinmiştir.

Tablo eksenleri alfabetik sıralandıktan sonra sonuç tablosu olarak yazdırılmıştır. Sonuç tablosunda 29870 gen ve 1155 hücre hattı bulunmaktadır.

**Metilasyon Verisi**

Metilasyon verisi için önceki aşamalarda oluşturulan düzenlenmiş CellMiner kaynak verisi kullanılarak hücre hattı ve gen isimlerinin eklenmesiyle tabloda genişletme işlemi yapılmıştır.

Verideki orijinal isimler düzenlenmiş halleriyle değiştirilmiştir.

GDSC için dört özellik veri tipi baz alarak oluşturulan en geniş hücre hattı ve gen listelerinde olup gen ifade verisi üzerinde olmayan isimler ilişkili eksenlere eklenmiştir. Yeni oluşan tablo boşluklarına 0 (metilasyon yok) değeri atanmıştır. Toplamda 10006 gen ve 75 hücre hattı tabloya eklenmiştir.

Tablo eksenleri alfabetik sıralandıktan sonra sonuç tablosu olarak yazdırılmıştır. Sonuç tablosunda 29870 gen ve 1155 hücre hattı bulunmaktadır.

**KSD Verisi**

KSD verisi için önceki aşamalarda oluşturulan düzenlenmiş kaynak veri (PICNIC algoritmasıyla üretilen) kullanılarak hücre hattı ve gen isimlerinin eklenmesiyle tabloda genişletme işlemi yapılmıştır.

Verideki orijinal isimler düzenlenmiş halleriyle değiştirilmiştir.

GDSC için dört özellik veri tipi baz alarak oluşturulan en geniş hücre hattı ve gen listelerinde olup gen ifade verisi üzerinde olmayan isimler ilişkili eksenlere eklenmiştir. Yeni oluşan tablo boşluklarına boşluk değeri (NaN, Not a Number) değeri atanmıştır. Toplamda 5368 gen ve 169 hücre hattı tabloya eklenmiştir.

Tablo eksenleri alfabetik sıralandıktan sonra sonuç tablosu olarak yazdırılmıştır. Sonuç tablosunda 29870 gen ve 1155 hücre hattı bulunmaktadır.

**İlaç Yanıtı Verisi**

Önceki aşamalarda hazırlanmış olan ilaç parmak izi tablosundaki parmak izi değerleri ilaç yanıtı tablosuna aktarılmıştır. Bu işlem için ilk olarak, düzenlen ilaç parmak izi tablosuna düzenlenmiş ilaç isimleri haritalandırılıp aktarıldı. Sonraki aşamada ise parmak izi değerleri ilaç yanıtı tablosuna düzenlenmiş ilaç isimleri üzerinden haritalandırılıp yeni bir sütun olarak eklenmiştir. Oluşturulan dosya kaydedilmiştir.

### 8.3.2. Hücre Hattı Özellik Vektörü Tablosunun Oluşturulması

Bu kısımda yukarıda bahsedilen aşamalarla oluşturulan dört özellik veri tipinin genişletilmiş veri setleri ve L1000 gen listesi kullanılarak tüm hücre hatlarının özellik vektörleri oluşturulmuştur.

İlk olarak, L1000 gen listesindeki isimler yukarıda bahsedilen isim düzenleme yöntemiyle istenilen formata getirilmiştir. Her özellik tipindeki isim listesi ve L1000 isim listesi karşılaştırıldı ve tüm L1000 genlerinin bu listelerde olduğu görüldü.

**Gen İfade Verisi**

Gen ifade verisi için gen bazında ortalamalar alınmış ve boşluk değeri içerenler için “NaN” değeri atanmıştır. Gen ismi ve ortalama değer sözlüğü oluşturulmuştur.

Gen ifade verisi için gen bazında doluluk oranını belirten tablo oluşturuldu. L1000 listesi kullanılarak doluluk oranı tablosu gen ismi sütunu üzerinden filtrelendi. Kalan 978 gen için yüzdelik sütunu üzerinden %87 ve üstü değere sahip genler için filtreleme yapıldı (%75 oran ile de bir değişiklik görülmedi).

Kalan gen listesi bir dosyaya yazdırıldı.

Tüm platformlarda aynı ortak gen ismi olması gerektiğinden, gen ifade verisi için platformların ortak gen ismi listesi kullanılarak bir filtreleme yapıldı. Sonuç olarak 897 gen ismi kalmıştır.

**KSD Verisi**

KSD verisi için gen bazında medyan değerleri alınmış ve boşluk değeri içerenler için “NaN” değeri atanmıştır. Gen ismi ve medyan değer sözlüğü oluşturulmuştur.

KSD verisi için gen bazında doluluk oranını belirten tablo oluşturuldu. L1000 listesi kullanılarak doluluk oranı tablosu gen ismi sütunu üzerinden filtrelendi. Kalan 978 gen için yüzdelik sütunu üzerinden %80 ve üstü değere sahip genler için filtreleme yapıldı (%75 oran ile de bir değişiklik görülmedi).

Kalan gen listesi bir dosyaya yazdırıldı.

Tüm platformlarda aynı ortak gen ismi olması gerektiğinden, KSD verisi için platformların ortak gen ismi listesi kullanılarak bir filtreleme yapıldı. Sonuç olarak 932 gen ismi kalmıştır.

**Mutasyon Verisi**

Mutasyon verisi için gen bazında 1 içeren tablo hücrelerinin sayısı hesaplanarak bir tablo oluşturuldu. Sayım değeri toplam hücre hattı sayısına bölünüp yeni bir sütunda doluluk yüzdesi yazdırıldı.

L1000 listesi kullanılarak doluluk oranı tablosu gen ismi sütunu üzerinden filtrelendi. Kalan 978 gen için yüzdelik sütunu üzerinden %1 ve üstü değere sahip genler için filtreleme yapıldı.

Kalan gen listesi bir dosyaya yazdırıldı.

Tüm platformlarda aynı ortak gen ismi olması gerektiğinden, mutasyon verisi için platformların ortak gen ismi listesi kullanılarak (son filtreleme iptal edilerek) bir filtreleme yapıldı. Sonuç olarak 963 gen ismi kalmıştır.

**Metilasyon Verisi**

Metilasyon verisi için gen bazında ortalama değerleri alınmış ve boşluk değeri içerenler için “NaN” değeri atanmıştır. Gen ismi ve ortalama değer sözlüğü oluşturulmuştur.

Metilasyon verisi için gen bazında doluluk oranını belirten tablo oluşturuldu. L1000 listesi kullanılarak doluluk oranı tablosu gen ismi sütunu üzerinden filtrelendi. Kalan 978 gen için yüzdelik sütunu üzerinden %75 ve üstü değere sahip genler için filtreleme yapıldı.

Kalan gen listesi bir dosyaya yazdırıldı.

Tüm platformlarda aynı ortak gen ismi olması gerektiğinden, metilasyon verisi için platformların ortak gen ismi listesi kullanılarak bir filtreleme yapıldı. Sonuç olarak 955 gen ismi kalmıştır.

**Hücre Hattı Özellik Vektörü Tablosu**

Yukarıda filtrelenen gen isimlerinin listeleri oluşturulup bu listelerle ilgili satırlar veri üzerinden elde edilmiştir.

Tüm tablolardaki indeks sütunu yeniden sıfır sayısından başlatılarak düzenlendi.

Sonuç tablosu 988 hücre hattı ve 3747 (897 + 963 + 955 + 932) gen ismi için boş değerli bir tablo oluşturuldu. Her bir hücre hattına ait değerler ilgili özellik verisinden çekilir ve tümü bir liste olarak birleştirilir. Bu 3747 uzunluğundaki bu liste sonuç tablosundaki ilgili hücre hattına ait satıra yerleştirilir. Eğer, herhangi bir özellik tipinden gelen veri içinde “NaN” etiketli değer bulunuyorsa bunun için farklı yöntemler izlenmiştir. Gen ifade verisi için, o gen ismine ait ortalama değer; mutasyon verisi için 0 değeri; metilasyon için ortalama değer; KSD için medyan değeri atanmıştır.

Sonuç tablosu, boşluk değeri içermediği kontrol edilip hücre hattı sütunu alfabetik sıralandıktan sonra bir dosyaya yazdırılıp kaydedilmiştir.

### 8.3.3. Çapraz Alan Analizi İçin CCLE Veri Tipleri Üzerinde Yapılan Ön İşlemler

#### Depmap KSD Tablosunun Düzenlenmesi

DepMap üzerinden CCLE'ye ait KSD verilerinden 22Q1 versiyonlu olan dosya (CCLE_gene_cn_22Q1_depmap.csv, https://ndownloader.figshare.com/files/34008428) indirilip üzerinde istenilen formata getirilmesi için ön işlemler yapılmıştır. 1 sayısı eklenip log2 tabanında hesaplanan sonucu kaynak tabloda olan veriler, önce karekök değeri alınıp sonra 1 çıkartılarak aynı tablo hücrelerine yazdırılmıştır.

Gen isimleri yine DepMap sitesinden alınan bilgilendirme tablosu (sample_info_22q1.csv, https://ndownloader.figshare.com/files/34008503) kullanılarak "ACH" kodlu hücre hatları bilinen isimleriyle değiştirilmiştir. Sütun başlarında yer alan gen isimleri ise parantez içi kodlardan temizlenip sadece isim kısmı bırakılmıştır. Her iki eksen de alfabetik sıralandıktan sonra transpozu alınan tablo yazdırılıp kaydedilmiştir.

#### Gen, Hücre Hattı ve İlaç İsimlerinin Düzenlenmesi

CCLE'deki 4 özellik veri tipi ve ilaç yanıtı verisi göz önüne alınarak gen ve hücre hatları isimleri üzerinde harf ve sayı dışındaki karakterlerin çıkarılıp küçük harfler ile yeniden düzenlenmesi işlemi, diğer platformlardaki isimlerle uyumlu olması için gerçekleştirilmiştir.

Düzenlenecek olan isim listeleri önceden oluşturulan düzenlenmiş kaynak veri yapıları üzerinden alınmıştır.

Hücre hatları için, "NCI" kodunu içeren isimlerden bu kod çıkarılmıştır.

Düzenlenmiş halleri aynı olan isimler olduğunda bu orijinal isimlerin sadece küçük harfli ve çıkarılacak karakterleri korunmuş halleri kullanılmıştır.

Sonuç olarak, tüm veriler kullanılarak en geniş haliyle 1800 hücre hattı ismi, 30509 gen ismi bulunmuştur ve tablo genişletme sırasında kullanılacak ana listeler böylece belirlenmiştir.

Tüm özellik tipleri ve ilaç yanıtı tablosu için oluşturulan sözlükler tablo halinde yazdırılıp kaydedilmiştir.

### 8.3.4. CCLE Omik Veri Tipi Tablolarının Genişletilmesi

#### Gen İfade Verisi

Gen ifade verisi için önceki aşamalarda oluşturulan düzenlenmiş CellMiner kaynak verisi kullanılarak hücre hattı ve gen isimlerinin eklenmesiyle tabloda genişletme işlemi yapılmıştır.

Verideki orijinal isimler düzenlenmiş halleriyle değiştirilmiştir.

CCLE için dört özellik veri tipi baz alarak oluşturulan en geniş hücre hattı ve gen listelerinde olup gen ifade verisi üzerinde olmayan isimler ilişkili eksenlere eklenmiştir. Yeni oluşan tablo boşluklarına boşluk değeri (NaN, Not a Number) değeri atanmıştır. Toplamda 10658 gen ve 712 hücre hattı tabloya eklenmiştir.

Tablo eksenleri alfabetik sıralandıktan sonra sonuç tablosu olarak yazdırılmıştır. Sonuç tablosunda 30509 gen ve 1800 hücre hattı bulunmaktadır.

#### Mutasyon Verisi

Mutasyon verisi için önceki aşamalarda oluşturulan düzenlenmiş kaynak veri kullanılarak hücre hattı ve gen isimlerinin eklenmesiyle tabloda genişletme işlemi yapılmıştır.

Verideki orijinal isimler düzenlenmiş halleriyle değiştirilmiştir.

CCLE için dört özellik veri tipi baz alarak oluşturulan en geniş hücre hattı ve gen listelerinde olup gen ifade verisi üzerinde olmayan isimler ilişkili eksenlere eklenmiştir. Yeni oluşan tablo boşluklarına 0 (mutasyon yok) değeri atanmıştır. Toplamda 11224 gen ve 230 hücre hattı tabloya eklenmiştir.

Tabloda oluşan “c1orf220” gen isminin birden çok olduğu görülünce ilk örnek tutulup diğeri silinmiştir.

Tablo eksenleri alfabetik sıralandıktan sonra sonuç tablosu olarak yazdırılmıştır. Sonuç tablosunda 30509 gen ve 1800 hücre hattı bulunmaktadır.

**Metilasyon Verisi**

Metilasyon verisi için önceki aşamalarda oluşturulan düzenlenmiş CellMiner kaynak verisi kullanılarak hücre hattı ve gen isimlerinin eklenmesiyle tabloda genişletme işlemi yapılmıştır.

Verideki orijinal isimler düzenlenmiş halleriyle değiştirilmiştir.

CCLE için dört özellik veri tipi baz alarak oluşturulan en geniş hücre hattı ve gen listelerinde olup gen ifade verisi üzerinde olmayan isimler ilişkili eksenlere eklenmiştir. Yeni oluşan tablo boşluklarına 0 (metilasyon yok) değeri atanmıştır. Toplamda 10629 gen ve 711 hücre hattı tabloya eklenmiştir.

Tablo eksenleri alfabetik sıralandıktan sonra sonuç tablosu olarak yazdırılmıştır. Sonuç tablosunda 30509 gen ve 1800 hücre hattı bulunmaktadır.

**KSD Verisi**

KSD verisi için önceki aşamalarda oluşturulan düzenlenmiş kaynak veri (DepMap) kullanılarak hücre hattı ve gen isimlerinin eklenmesiyle tabloda genişletme işlemi yapılmıştır.

Verideki orijinal isimler düzenlenmiş halleriyle değiştirilmiştir.

CCLE için dört özellik veri tipi baz alarak oluşturulan en geniş hücre hattı ve gen listelerinde olup gen ifade verisi üzerinde olmayan isimler ilişkili eksenlere eklenmiştir. Yeni oluşan tablo boşluklarına boşluk değeri (NaN, Not a Number) değeri atanmıştır. Toplamda 5141 gen ve 46 hücre hattı tabloya eklenmiştir.

Tablo eksenleri alfabetik sıralandıktan sonra sonuç tablosu olarak yazdırılmıştır. Sonuç tablosunda 30509 gen ve 1800 hücre hattı bulunmaktadır.

**İlaç Yanıtı Verisi**

Önceki aşamalarda hazırlanmış olan ilaç parmak izi tablosundaki parmak izi değerleri ilaç yanıtı tablosuna aktarılmıştır. Bu işlem için ilk olarak, düzenlen ilaç parmak izi tablosuna düzenlenmiş ilaç isimleri haritalandırılıp aktarıldı. Sonraki aşamada ise parmak izi değerleri ilaç yanıtı tablosuna düzenlenmiş ilaç isimleri üzerinden haritalandırılıp yeni bir sütun olarak eklenmiştir. Oluşturulan dosya kaydedilmiştir.

### 8.3.5. CCLE verisindeki ilaç isimlerine karşılık SMILES dizilerinin bulunması

Pandas ve NumPy kütüphaneleri hazır hale getirilip CCLE ilaç yanıtı verisi bir değişkene aktarılmıştır. İlaç isimleri bir listeye aktarılmıştır. Ardından, bir sözlük oluşturularak isimler için boş değerler atanmıştır. İlaç yanıtı verisinde SMILES sütunu oluşturulup boş değerler atanmıştır.

Pubchempy kütüphanesi hazır hale getirildikten sonra sözlükteki her ilaç için bir döngü içinde sorgu oluşturulup ilaca ait standart SMILES dizisi bir değer olarak sözlüğe eklenmiştir.

Yapılandırılan sözlük bir tablo haline getirilmiştir. Oluşan tablo bir dosyaya yazdırılarak kaydedilmiştir.

### 8.3.6. CCLE verisindeki ilaçlar için parmak izi oluşturulması

Pandas ve NumPy kütüphaneleri hazırlandıktan sonra SMILES karşılıkları bulunan CCLE ilaçlarının olduğu " CCLE_mapped_drugs_smiles_df_v1.txt" dosyasındaki tablo bir değişkene atanmıştır.

Sys ve RDkit kütüphaneleri kullanılmaya uygun hale getirilip tabloya ECFP4 isimli boş değerli sütun eklenmiştir.

İlaçların SMILES dizilerinin bulunduğu sütundaki değerlerin teker teker işlenerek ECFP4 sütununa karşılıklarının yazıldığı bir döngü oluşturulmuştur. Bu döngüde, SMILES dizileri GetMorganFingerprintAsBitVect metodu ve radius (2 değeri atanmıştır), nBits (1024 değeri atanmıştır) parametreleri yardımıyla ilaç parmak izlerine dönüştürülmüştür. Parmak izi oluşturma işlemi iki adımda gerçekleştiği ve nBits parametresi 1024 olarak belirlendiği için parmak izleri, ECFP_4 formatında ve 1024 elemanlı (bit) olarak üretilmiştir.

Sonuç tablosu bir dosyaya yazdırılıp kaydedilmiştir.

### 8.3.7. CCLE Hücre Hattı Özellik Vektörü Tablosunun Oluşturulması

Bu kısımda yukarıda bahsedilen aşamalarla oluşturulan dört özellik veri tipinin genişletilmiş veri setleri ve L1000 gen listesi kullanılarak tüm hücre hatlarının özellik vektörleri oluşturulmuştur.

İlk olarak, L1000 gen listesindeki isimler yukarıda bahsedilen isim düzenleme yöntemiyle istenilen formata getirilmiştir. Her özellik tipindeki isim listesi ve L1000 isim listesi karşılaştırıldı ve 970 L1000 geninin bu listelerde olduğu görüldü.

#### Gen İfade Verisi

Gen ifade verisi için gen bazında ortalamalar alınmış ve boşluk değeri içerenler için “NaN” değeri atanmıştır. Gen ismi ve ortalama değer sözlüğü oluşturulmuştur.

Gen ifade verisi için gen bazında doluluk oranını belirten tablo oluşturuldu. L1000 listesi kullanılarak doluluk oranı tablosu gen ismi sütunu üzerinden filtrelendi. Kalan 970 gen için yüzdelik sütunu üzerinden %57 ve üstü değere sahip genler için filtreleme yapıldı (%75 oran ile de bir değişiklik görülmedi).

Kalan gen listesi bir dosyaya yazdırıldı.

Tüm platformlarda aynı ortak gen ismi olması gerektiğinden, gen ifade verisi için platformların ortak gen ismi listesi kullanılarak bir filtreleme yapıldı. Sonuç olarak 897 gen ismi kalmıştır.

**KSD Verisi**

KSD verisi için gen bazında medyan değerleri alınmış ve boşluk değeri içerenler için “NaN” değeri atanmıştır. Gen ismi ve medyan değer sözlüğü oluşturulmuştur.

KSD verisi için gen bazında doluluk oranını belirten tablo oluşturuldu. L1000 listesi kullanılarak doluluk oranı tablosu gen ismi sütunu üzerinden filtrelendi. Kalan 970 gen için yüzdelik sütunu üzerinden %80 ve üstü değere sahip genler için filtreleme yapıldı.

Kalan gen listesi bir dosyaya yazdırıldı.

Tüm platformlarda aynı ortak gen ismi olması gerektiğinden, KSD verisi için platformların ortak gen ismi listesi kullanılarak bir filtreleme yapıldı. Sonuç olarak 932 gen ismi kalmıştır.

**Mutasyon Verisi**

Mutasyon verisi için gen bazında 1 içeren tablo hücrelerinin sayısı hesaplanarak bir tablo oluşturuldu. Sayım değeri toplam hücre hattı sayısına bölünüp yeni bir sütunda doluluk yüzdesi yazdırıldı.

L1000 listesi kullanılarak doluluk oranı tablosu gen ismi sütunu üzerinden filtrelendi. Kalan 840 gen için yüzdelik sütunu üzerinden %1 ve üstü değere sahip genler için filtreleme yapıldı.

Kalan gen listesi bir dosyaya yazdırıldı.

Tüm platformlarda aynı ortak gen ismi olması gerektiğinden, mutasyon verisi için platformların ortak gen ismi listesi kullanılarak (son filtreleme iptal edilerek) bir filtreleme yapıldı. Sonuç olarak 963 gen ismi kalmıştır.

**Metilasyon Verisi**

Metilasyon verisi için gen bazında ortalama değerleri alınmış ve boşluk değeri içerenler için “NaN” değeri atanmıştır. Gen ismi ve ortalama değer sözlüğü oluşturulmuştur.

Metilasyon verisi için gen bazında doluluk oranını belirten tablo oluşturuldu. L1000 listesi kullanılarak doluluk oranı tablosu gen ismi sütunu üzerinden filtrelendi. Kalan 978 gen için yüzdelik sütunu üzerinden %75 ve üstü değere sahip genler için filtreleme yapıldı.

Kalan gen listesi bir dosyaya yazdırıldı.

Tüm platformlarda aynı ortak gen ismi olması gerektiğinden, metilasyon verisi için platformların ortak gen ismi listesi kullanılarak bir filtreleme yapıldı. Sonuç olarak 955 gen ismi kalmıştır.

**Hücre Hattı Özellik Vektörü Tablosu**

Yukarıda filtrelenen gen isimlerinin listeleri oluşturulup bu listelerle ilgili satırlar veri üzerinden elde edilmiştir.

Tüm tablolardaki indeks sütunu yeniden sıfır sayısından başlatılarak düzenlendi.

Sonuç tablosu 503 hücre hattı ve 3747 (897 + 963 + 955 + 932) gen ismi için boş değerli bir tablo oluşturuldu. Her bir hücre hattına ait değerler ilgili özellik verisinden çekilir ve tümü bir liste olarak birleştirilir. Bu 3747 uzunluğundaki bu liste sonuç tablosundaki ilgili hücre hattına ait satıra yerleştirilir. Eğer, herhangi bir özellik tipinden gelen veri içinde “NaN” etiketli değer bulunuyorsa bunun için farklı yöntemler izlenmiştir. Gen ifade verisi için, o gen ismine ait ortalama değer; mutasyon verisi için 0 değeri; metilasyon için ortalama değer; KSD için medyan değeri atanmıştır.

Sonuç tablosu, boşluk değeri içermediği kontrol edilip hücre hattı sütunu alfabetik sıralandıktan sonra bir dosyaya yazdırılıp kaydedilmiştir.

### 8.3.8. GDSC – CCLE Çapraz Alan Analizindeki Modeller İçin Gereken Veri Setlerinin Oluşturulması

GDSC ve CCLE için çapraz alan analizinde kullanılmak üzere oluşturulan 3747 vektör uzunluğunda olan hücre hattı özellik tabloları ile beraber farklı modelleme senaryoları oluşturulmuştur. Öncelikle, kullanılacak olan bu iki veri arasındaki hücre hatlarının ortaklıkları incelenmiştir. GDSC ilaç isimlerinden "nutlin3a ", “nvptae684” sırasıyla “nutlin3”, “tae684” ile değiştirilmiştir. CCLE ilaç yanıtı tablosundaki hücre hatlarından “nihovcar3” ismi “ovcar3” ile değiştirilmiştir. İki platforma ait olan ilaç yanıtı verileri üzerinden de hücre hattı ve ilaç ortaklıkları incelenmiştir. Ortak olan 376 hücre hattı ve 13 ilaç ismi olduğu görülmüştür.

#### Analizde Kullanılacak Olan İlaç Yanıtı Verilerinin Düzenlenmesi

GDSC’de olup CCLE’de olmayan 612 hücre hattı için ilaç yanıtı verisi üzerinden bu hatlara ait satırları almak üzere filtreleme yapılarak 327231 satırlı tablo senaryo 1 için oluşturuldu ve sonuç bir dosyaya yazdırıldı. Aynı şekilde, GDSC’de bulunup CCLE’de olmayan 391 ilaç için ilaç yanıtı verisi üzerinden bu ilaçlara ait satırları almak üzere filtreleme yapılarak 203934 satırlı tablo senaryo 2 için oluşturuldu ve sonuç bir dosyaya yazdırıldı.

**Tablo 8.8.** GDSC – CCLE arasında oluşturulan çapraz alan analizi için oluşturulan senaryolarda kullanılacak olan verilerin düzenlenme planları.

| | **GDSC** | | **CCLE** | |
|---|---|---|---|---|
| **Senaryo** | **Hücre Hattı Özellik Vektörü Dosyası** | **İlaç Yanıtı Vektör Dosyası** | **Hücre Hattı Özellik Vektörü Dosyası** | **İlaç Yanıtı Dosyası** |
| **1** | Ortak olmayan hücre hattı | Ortak olmayan hücre hattı | Tümü | Tümü |
| **2** | Ortak olmayan ilaçlara ait hücre hatları | Ortak olmayan ilaç | Tümü | Tümü |
| **3** | Tümü | Tümü | Ortak hücre hattı | Ortak hücre hattı |
| **4** | Tümü | Tümü | Ortak olmayan hücre hattı | Ortak olmayan hücre hattı |
| **5** | Tümü | Tümü | Ortak ilaçlara ait hücre hatları | Ortak ilaç |
| **6** | Tümü | Tümü | Ortak olmayan ilaçlara ait hücre hatları | Ortak olmayan ilaç |
| **7** | Tümü | Tümü | Tümü | Tümü |

Senaryo 1 ve 2 için CCLE'nin tüm ilaç yanıtı verisi kullanılmıştır. Senaryo 3, 4, 5, 6, 7 için GDSC'nin tüm ilaç yanıtı verisi kullanılmıştır.

Ortak hücre hatları, CCLE – GDSC farkı hücre hatları, Ortak ilaçlar, CCLE – GDSC farkı ilaç listeleri için ilaç yanıtı tablosundan sırasıyla yapılan filtrelemelerle 8666, 3004, 6265, 5405 satırlı tablolar oluşturulup ayrı ayrı dosyalara yine sırasıyla senaryo 3, 4, 5, 6 için kaydedilmiştir.

**Analizde Kullanılacak Olan Hücre Hattı Özellik Vektörü Verilerinin Düzenlenmesi**

GDSC – CCLE farkı hücre hattı ve senaryo 2'deki hücre hattı listesi kullanılarak GDSC hücre hattı özellik vektörü dosyasından sırasıyla yapılan filtrelemelerle 612, 988 satırlı tablolar oluşturulup ayrı ayrı dosyalara yine sırasıyla senaryo 1, 2 için kaydedilmiştir.

Senaryo 1 ve 2 için CCLE'nin tüm hücre hattı vektör verisi kullanılmıştır. Senaryo 3, 4, 5, 6, 7 için GDSC'nin tüm hücre hattı vektör verisi kullanılmıştır.

CCLE için oluşturulan senaryo 3, 4, 5, 6 ilaç yanıtı tablolarındaki hücre hattı listeleri çıkarılarak CCLE hücre hattı özellik vektöründeki hücre hattı sütunu üzerinden filtreleme yapılarak sırasıyla 376, 127, 503, 503 satırlı tablolar elde edilmiş ve senaryo 3 ve 4 için sonuçlar dosyalara yazdırılıp kaydedilmiştir. Senaryo 5 ve 6 için ise yine tablodaki tüm hücre hatları elde edildiğinden farklı bir dosyaya yazdırmaya gerek görülmemiştir.

### 8.3.9. Çapraz Alan Analizi İçin NCI-60 Veri Tipleri Üzerinde Yapılan Ön İşlemler

**Gen İfade Kaynak Verisinin Düzenlenmesi**

Gen ifade verisi için CellMinerCDB web sitesinden (https://discover.nci.nih.gov/rsconnect/cellminercdb/) indirilen mRNA ifade dosyası (data_NCI-60_xai.zip) kullanılmıştır. Tablo düzenlenmesinde hücre hattı ismi olmayan sütunlar çıkartılmıştır. Hücre hattı isimlerinin ön kısımlarında bulunan kodlar silinerek normal isimler tekrar tabloya yazdırılmıştır. Tablo eksenleri tekrar alfabetik sıralanıp yazdırılmıştır ve bir dosyaya kaydedilmiştir.

**Mutasyon Kaynak Verisinin Düzenlenmesi**

"A survey and systematic assessment of computational methods for drug response prediction" başlıklı makaleye ait ek dosyalarda

(https://github.com/Jinyu2019/Suppl-data-BBpaper/blob/master/NCI60_DATASET_S27-S39.zip) bulunan mutasyon verisi (Table_S31_NCI60_Mutation.csv) NCI-60 için CellMiner verisi yerine sadece ikili (0, 1) değer içeren yapıda olduğu için tercih edilmiştir.

Bu verinin düzenlenmesi için öncelikle tablonun transpozu alınarak istenilen eksenler elde edilmiştir. Hücre hattı isimlerinde bulunan kod kısımları silinerek çıkarılmıştır. Gen ve hücre hattı listelerindeki isimler harf ve sayı dışındaki karakterler silinip küçük harf kullanılarak tabloya yeniden yazdırılmıştır. Eksenleri yeniden alfabetik sıralanan tablo yazdırılıp bir dosyaya kaydedilmiştir.

**KSD Kaynak Verisinin Düzenlenmesi**

KSD verisi için CellMiner web sitesinden elde edilen seti olan "DNA__aCGH_Agilent_44K_Copy_number_estimate.zip" dosyası kullanılmıştır. İstenilen formatla ilgisiz olan satırlar ve sütunlar çıkarılmıştır. Hücre hatları isimlerinin önündeki kodlar silinerek asıl isimleri korunarak tabloya tekrar yazdırılmıştır. Gen isimlerinde "-" değerine sahip olan gen satırları çıkarılmıştır. Eksenleri alfabetik olarak sıralanan tablo yazdırılıp bir dosyaya kaydedilmiştir.

### 8.3.10. NCI-60 Gen, Hücre Hattı ve İlaç İsimlerinin Düzenlenmesi

NCI-60'deki 4 özellik veri tipi ve ilaç yanıtı verisi göz önüne alınarak gen ve hücre hatları isimleri üzerinde harf ve sayı dışındaki karakterlerin çıkarılıp küçük harfler ile yeniden düzenlenmesi işlemi, diğer platformlardaki isimlerle uyumlu olması için gerçekleştirilmiştir.

Düzenlenecek olan isim listeleri önceden oluşturulan düzenlenmiş kaynak veri yapıları üzerinden alınmıştır.

Düzenlenmiş halleri aynı olan isimler olduğunda bu orijinal isimlerin sadece küçük harfli ve çıkarılacak karakterleri korunmuş halleri kullanılmıştır.

Sonuç olarak, tüm veriler kullanılarak en geniş haliyle 60 hücre hattı ismi, 25647 gen ismi bulunmuştur ve tablo genişletme sırasında kullanılacak ana listeler böylece belirlenmiştir.

Tüm özellik tipleri ve ilaç yanıtı tablosu için oluşturulan sözlükler tablo halinde yazdırılıp kaydedilmiştir.

### 8.3.11. NCI-60 Omik Veri Tipi Tablolarının Genişletilmesi

#### Gen İfade Verisi

Gen ifade verisi için önceki aşamalarda oluşturulan düzenlenmiş CellMiner kaynak verisi kullanılarak hücre hattı ve gen isimlerinin eklenmesiyle tabloda genişletme işlemi yapılmıştır.

Verideki orijinal isimler düzenlenmiş halleriyle değiştirilmiştir.

NCI-60 için dört özellik veri tipi baz alarak oluşturulan en geniş hücre hattı ve gen listelerinde olup gen ifade verisi üzerinde olmayan isimler ilişkili eksenlere eklenmiştir. Yeni oluşan tablo boşluklarına boşluk değeri (NaN, Not a Number) değeri atanmıştır. Toplamda 2588 gen tabloya eklenmiştir.

Tablo eksenleri alfabetik sıralandıktan sonra sonuç tablosu olarak yazdırılmıştır. Sonuç tablosunda 25647 gen ve 60 hücre hattı bulunmaktadır.

#### Mutasyon Verisi

Mutasyon verisi için yukarıda bahsedilen makaleye ait mutasyon verisinin düzenlenmiş hali kaynak veri olarak kullanılarak hücre hattı ve gen isimlerinin eklenmesiyle veride genişletme işlemi yapılmıştır.

Verideki orijinal isimler düzenlenmiş halleriyle değiştirilmiştir.

NCI-60 için dört özellik veri tipi baz alarak oluşturulan en geniş hücre hattı ve gen listelerinde olup gen ifade verisi üzerinde olmayan isimler ilişkili eksenlere eklenmiştir. Yeni oluşan tablo boşluklarına 0 (mutasyon yok) değeri atanmıştır. Toplamda 25204 gen ve 1 hücre hattı tabloya eklenmiştir.

Tablo eksenleri alfabetik sıralandıktan sonra sonuç tablosu olarak yazdırılmıştır. Sonuç tablosunda 25647 gen ve 60 hücre hattı bulunmaktadır.

**Metilasyon Verisi**

Metilasyon verisi için önceki aşamalarda oluşturulan düzenlenmiş CellMiner kaynak verisi kullanılarak hücre hattı ve gen isimlerinin eklenmesiyle tabloda genişletme işlemi yapılmıştır.

Verideki orijinal isimler düzenlenmiş halleriyle değiştirilmiştir.

NCI-60 için dört özellik veri tipi baz alarak oluşturulan en geniş hücre hattı ve gen listelerinde olup gen ifade verisi üzerinde olmayan isimler ilişkili eksenlere eklenmiştir. Yeni oluşan tablo boşluklarına 0 (metilasyon yok) değeri atanmıştır. Toplamda 8095 gen tabloya eklenmiştir.

Tablo eksenleri alfabetik sıralandıktan sonra sonuç tablosu olarak yazdırılmıştır. Sonuç tablosunda 25647 gen ve 60 hücre hattı bulunmaktadır.

**KSD Verisi**

KSD verisi için önceki aşamalarda oluşturulan düzenlenmiş CellMiner kaynak verisi kullanılarak hücre hattı ve gen isimlerinin eklenmesiyle tabloda genişletme işlemi yapılmıştır.

Verideki orijinal isimler düzenlenmiş halleriyle değiştirilmiştir.

NCI-60 için dört özellik veri tipi baz alarak oluşturulan en geniş hücre hattı ve gen listelerinde olup gen ifade verisi üzerinde olmayan isimler ilişkili eksenlere eklenmiştir.

Yeni oluşan tablo boşluklarına boşluk değeri (NaN, Not a Number) değeri atanmıştır. Toplamda 5800 gen tabloya eklenmiştir.

Tablo eksenleri alfabetik sıralandıktan sonra sonuç tablosu olarak yazdırılmıştır. Sonuç tablosunda 25647 gen ve 60 hücre hattı bulunmaktadır.

### 8.3.12. NCI-60 İlaç Yanıtı Veri Tipi Düzenlemeleri

#### İlaç Yanıtı Verisi

NCI-60 verisi R programlama dilinde bulunan Rcellminer paketi içindeki veri yapılarından elde edilmiştir. Bunun için öncelikle rcellminer ve rcellminerData paketleri hazır hale getirilmiştir. Ardından, negatif logaritmalı GI50 değerlerinin bir tablo olarak elde edilmesi için https://rdrr.io/bioc/rcellminer/f/vignettes/rcellminerUsage.Rmd sayfasında yer alan yönergeler izlenmiştir. Sonuç tablosu bir dosyaya yazdırılarak kaydedilmiştir.

İlaç yanıtı tablosunun istenilen formata getirilmesi Python'daki Pandas kütüphanesi yardımıyla yapılmıştır. Tablodaki ilaç kimliklerine karşılık gelen isimlerin bulunması için NCI-60_drug_name_nsc_id_smiles_df_v1.txt, hemnames_Aug2013.zip, DTP_NCI60_ZSCORE.xlsx dosyalarından yararlanılmıştır. 9 kimlik için PubChem üzerinden sorgulama yapılarak karşılıklar elle atanmıştır. İsim karşılıkları tabloya yansıtıldıktan sonra isim sütununda aynı değeri içeren satırlardan sadece ilk örnekler tutulmuştur. Tablo, isim ve kimlik sütunlarına göre alfabetik sıralanmıştır. Hücre hattı isimlerindeki kodlar ve ayraç kaldırılmıştır ve düzenlemeler için bir sözlük oluşturulmuştur.

İlaç ismi, kimliği, hücre hattı ismi ve pGI50 sütunlarıyla yeni bir tablo oluşturulmuştur. Bu tablo için ana tablodaki bu sütun değerlerine karşılık gelecek tüm listeler, ilaç yanıtı değerlerinin tümüne karşılık gelen 64080 sayısı uzunluğunda olacak şekilde çarpılarak uzatılmıştır. Bu listeler yeni bir tablo üzerinde dikey olarak birleştirilmiştir. İlaç yanıtı sütununa ise boş değerler atanmıştır. Ana tablodan bir döngü

yardımıyla pGI50 değerleri 60'ar listeler şeklinde yeni tabloya aktarılmıştır. Düzenlenmiş hücre hattı isimleri de tabloya tekrardan üçüncü sütun olarak eklenmiş ilk hale sahip hücre hattı sütunu silinmiştir. Tabloya, ek olarak ilaç ve hücre hattı isimlerinin küçük harfle yazdırılmış, harf ve sayı dışındaki karakterleri silinmiş halleri yeni iki sütun olarak eklenmiştir. Son olarak, NCI-60_drug_name_SMILES_and_ECFP4_v3.txt dosyasından elde edilen ilaç parmak izi değerleri tabloya ilaç kimliği sütunu üzerinden yansıtılarak eklenmiştir. Tabloda düzenlenmiş haldeki ilaç ve hücre hattı sütunları için, aynı hücre hattı – ilaç ismi çiftleri silinmiştir. İlaç parmak izi sütununda boşluk değeri olan satırlar tablodan silinmiştir. Sonuç tablosu bir dosyaya yazdırılarak kaydedilmiştir.

**NCI-60 Verisindeki İlaç İsimlerine Karşılık SMILES Dizilerinin Bulunması**

NCI-60 ilaçlarına ait kimlik değerlerinin bulunduğu NSC_CAS_Sept2013.csv dosyası Pandas ve NumPy kütüphaneleri hazırlandıktan sonra bir değişkene aktarılmıştır. Tablodaki kimlik ve Kimyasal Özetler Hizmeti (Chemical Abstracts Service, CAS) numaraları bir sözlükte toplanmıştır.

İlaç kimliklerini sorgulayarak ilaç isimlerini elde etme işlemi uygulanmıştır. Bunun için, ast ve urllib kütüphaneleri hazır duruma getirilip oluşturulan sözlük üzerinden bir döngü oluşturulmuştur. Bu döngü yardımıyla, ChemIDPlus API (Application Programming Interface, uygulama programlama arayüzü) https://chem.nlm.nih.gov/api/data/rn/equals/ bağlantısı üzerinden ilaç kimliği sorgulaması yapılıp bulunan isim bir sözlüğe kimliklerle beraber kaydedilmiştir.

İlaç yanıtı verisi bir değişkene atanıp ilaç kimliği değerleri bir listeye aktarılmıştır. Liste 55600 uzunluğunda olduğu için sorgulamalar parçalara ayrılmıştır. Bunu uygulamak için, liste 10 parçaya bölünmüştür ve yeni bir liste olarak kaydedilmiştir.

Requests kütüphanesi hazırlanıp bölümlenen listenin içindeki her listenin sorgulama değeri olduğu bir döngü oluşturulmuştur. Her 49 dosyada bir olacak şekilde

60 saniye ara verilip döngüye devam edilmiştir. Her döngü sonunda sonuçlar bir dosyaya yazdırılıp farklı bir isimle kaydedilmiştir.

Kaydedilen dosyaların düzenlenmesi için bir döngü oluşturulup yeni dosyalar oluşturulmuştur. Kaydedilen dosyalardaki her bir satırda yer alan isimler öncelikle birbirinden ayrılmıştır. Ardından, yeni bir satır olarak yeni dosyaya yazdırılmıştır.

Düzenlenen dosyaların yatay olarak birleştirilmesi için bir döngü oluşturulmuştur. Her döngüde bir dosyaya ait tablo öncekine yatay olarak eklenerek kaydedilmiştir. Sonuç tablosu bir dosyaya yazdırılıp kaydedilmiştir.

Birleştirilen tablodaki birbirinin kopyası olan satırlardan sadece ilkleri tutulup diğerleri silinmiştir. Sonrasında, tablo ilaç kimliği sütununa göre gruplanmıştır. Kimlik ve isim sütunlarıyla bir sözlük oluşturulmuştur. Sözlükteki her değer uzunluklarına göre düzenlenip aynı sözlüğe bir liste olarak kaydedilmiştir.

PubChem verisine erişim sağlayan web arayüzü kullanılarak ilaç isimleri sorgulanıp standart SMILES dizileri bulunmuştur. Bunun için, PUG-REST (Power User Gateway - Representational State Transfer, Yetkili Kullanıcı Ağ Geçidi - Temsili Durum Transferi) üzerinden sözlükteki her ilaç ismi için bir sorgulama döngüsü oluşturulmuştur. Bulunan SMILES değerleri bir sözlükte isimlerle beraber kaydedilmiştir.

İlaç isimleri ve kimliklerinin iki sütun halinde bulunduğu önceden oluşturulan bir sözlüğün yardımıyla oluşturulmuştur. Elde edilen SMILES dizileri bu tabloya ilaç ismi sütunu üzerinden sözlük yardımıyla yansıtılarak yeni sütun olarak eklenmiştir.

Tablodaki boş yerlere “NaN” değeri atanmıştır. Tablo kimlik ve SMILES sütununa göre alfabetik sıralanmıştır. Birden çok kimlik değeri olan ilaçlardan ilki tutulup diğeri silinmiştir. Tablo, bir dosyaya yazdırılarak NCI-60_drug_name_nsc_id_smiles_df_v2.txt adıyla kaydedilmiştir.

**NCI-60 Verisindeki İlaçlar İçin Parmak İzi Değerlerinin Oluşturulması**

Pandas ve NumPy kütüphaneleri hazırlandıktan sonra NCI-60_drug_name_nsc_id_smiles_df_v2.txt dosyasındaki tablo bir değişkene atanmıştır.

Sys ve RDkit kütüphaneleri kullanılmaya uygun hale getirilip tabloya ECFP4 isimli boş değerli sütun eklenmiştir.

İlaçların SMILES dizilerinin bulunduğu sütunldaki değerlerin teker teker işlenerek ECFP4 sütununa karşılıklarının yazıldığı bir döngü oluşturulmuştur. Bu döngüde, SMILES dizileri GetMorganFingerprintAsBitVect metodu ve *radius* (2 değeri atanmıştır), nBits (1024 değeri atanmıştır) parametreleri yardımıyla ilaç parmak izlerine dönüştürülmüştür. Parmak izi oluşturma işlemi iki adımda gerçekleştiği ve nBits parametresi 1024 olarak belirlendiği için parmak izleri, ECFP_4 formatında ve 1024 elemanlı (bit) olarak üretilmiştir. Tablodaki boşluk değerine sahip satırlar silinerek NCI-60_drug_name_SMILES_and_ECFP4_nan_dropped_v2.txt adıyla kaydedilmiştir.

Önceki adımlarda oluşturulan ilaç yanıtı verisi bir değişkene aktarıldıktan sonra, ilaç kimlikleri bir listeye alınmıştır. Ardından, yukarıda oluşturulan son dosyadaki kimlikler de listelenip diğer listeyle karşılaştırılmıştır. Sadece ilaç yanıtı verisinde bulunan kimliklere ait satırlar veri içinden filtrelenmiştir. Kalan satırlardaki ilaç isimleri listelenmiştir. Bu listenin kopyası oluşturulup her ilacın önüne “nsc” kodu getirilerek başka bir liste daha yaratılmıştır.

Pathlib, Urllib, Requests kütüphaneleri hazır hale getirilip iki ilaç ismi listesi için de ayrı döngü oluşturulup PubChem web sitesi arayüzünden isimlere ait standart SMILES dizileri elde edilmiştir. Sonuç değerleri bir listede toplanmıştır. Sonrasında, sonuç sözlükteki ilaç isimlerinden “nsc” ekleri çıkarılarak başka bir listeye SMILES değerleriyle kaydedilmiştir. Diğer ek içermeyen isim sözlüğü de köken aldığı isim sözlüğü ile karşılaştırılıp sözlük değerleri yeni bir sözlüğe aktarılmıştır. Yeni oluşturulan sözlüklerin anahtar değerleri kendi aralarında karşılaştırılıp sonuçlar yeni listelere aktarılmıştır. Son

adım olarak, “nsc” ekini içeren isimlere ait sözlük diğer sözlük değerleriyle güncellenmiştir. Ana kimlik listesinde bulunan ancak SMILES dizisi bulunamayan “271674” kimlikli ilaç için PubChem sitesinden bulunan SMILES değeri direkt olarak atanmıştır.

SMILES değerlerini içeren ana tablo ilaç yanıtı verisiyle ortak olan ilaçlar ile kimlik sütunu üzerinden filtrelenmiştir. Diğer yandan, ilaç yanıtı verisinden filtrelenmiş değerlerle oluşturulan ilaç isim ve kimlik sözlüğü bir tablo haline getirilmiştir. Bu tablo üzerine kimlik sütunu üzerinden sonuç sözlüğü yansıtılarak SMILES sütunu oluşturulmuştur. Tablodaki boş değere sahip satırlar silinmiştir.

Sys, RDKit kütüphaneleri hazır hale getirilip önceki adımlarda uygulanan ilaç parmak izi çıkarma işlemi uygulanmıştır. Sonuçlar tabloya son sütun olarak ECFP4 ismiyle eklenmiştir.

Önceden filtrelenen SMILES değerlerini içeren tablo ve son olarak düzenlenen tablo yatay eksende birleştirilmiştir. Elde edilen veri bir dosyaya yazdırılarak kaydedilmiştir.

### 8.3.13. Hücre Hattı Özellik Vektörü Tablosunun Oluşturulması

Bu kısımda yukarıda bahsedilen aşamalarla oluşturulan dört özellik veri tipinin genişletilmiş veri setleri ve L1000 gen listesi kullanılarak tüm hücre hatlarının özellik vektörleri oluşturulmuştur.

İlk olarak, L1000 gen listesindeki isimler yukarıda bahsedilen isim düzenleme yöntemiyle istenilen formata getirilmiştir. Her özellik tipindeki isim listesi ve L1000 isim listesi karşılaştırıldı ve 963 L1000 geninin bu listelerde olduğu görüldü.

**Gen İfade Verisi**

Gen ifade verisi için gen bazında ortalamalar alınmış ve boşluk değeri içerenler için “NaN” değeri atanmıştır. Gen ismi ve ortalama değer sözlüğü oluşturulmuştur.

Gen ifade verisi için gen bazında doluluk oranını belirten tablo oluşturuldu. L1000 listesi kullanılarak doluluk oranı tablosu gen ismi sütunu üzerinden filtrelendi. Kalan 951 gen için yüzdelik sütunu üzerinden %75 ve üstü değere sahip genler için filtreleme yapıldı.

Kalan gen listesi bir dosyaya yazdırıldı.

Tüm platformlarda aynı ortak gen ismi olması gerektiğinden, gen ifade verisi için platformların ortak gen ismi listesi kullanılarak bir filtreleme yapıldı. Sonuç olarak 897 gen ismi kalmıştır.

**KSD Verisi**

KSD verisi için gen bazında medyan değerleri alınmış ve boşluk değeri içerenler için “NaN” değeri atanmıştır. Gen ismi ve medyan değer sözlüğü oluşturulmuştur.

KSD verisi için gen bazında doluluk oranını belirten tablo oluşturuldu. L1000 listesi kullanılarak doluluk oranı tablosu gen ismi sütunu üzerinden filtrelendi. Kalan 963 gen için yüzdelik sütunu üzerinden %75 ve üstü değere sahip genler için filtreleme yapıldı.

Kalan gen listesi bir dosyaya yazdırıldı.

Tüm platformlarda aynı ortak gen ismi olması gerektiğinden, KSD verisi için platformların ortak gen ismi listesi kullanılarak bir filtreleme yapıldı. Sonuç olarak 932 gen ismi kalmıştır.

**Mutasyon Verisi**

Mutasyon verisi için gen bazında 1 içeren tablo hücrelerinin sayısı hesaplanarak bir tablo oluşturuldu. Sayım değeri toplam hücre hattı sayısına bölünüp yeni bir sütunda doluluk yüzdesi yazdırıldı.

L1000 listesi kullanılarak doluluk oranı tablosu gen ismi sütunu üzerinden filtrelendi. Kalan 963 gen için yüzdelik sütunu üzerinden %1 ve üstü değere sahip genler için filtreleme yapıldı.

Kalan gen listesi bir dosyaya yazdırıldı.

Tüm platformlarda aynı ortak gen ismi olması gerektiğinden, mutasyon verisi için platformların ortak gen ismi listesi kullanılarak (son filtreleme iptal edilerek) bir filtreleme yapıldı. Sonuç olarak 963 gen ismi kalmıştır.

**Metilasyon Verisi**

Metilasyon verisi için gen bazında ortalama değerleri alınmış ve boşluk değeri içerenler için “NaN” değeri atanmıştır. Gen ismi ve ortalama değer sözlüğü oluşturulmuştur.

Metilasyon verisi için gen bazında doluluk oranını belirten tablo oluşturuldu. L1000 listesi kullanılarak doluluk oranı tablosu gen ismi sütunu üzerinden filtrelendi. Kalan 963 gen için yüzdelik sütunu üzerinden %75 ve üstü değere sahip genler için filtreleme yapıldı.

Kalan gen listesi bir dosyaya yazdırıldı.

Tüm platformlarda aynı ortak gen ismi olması gerektiğinden, metilasyon verisi için platformların ortak gen ismi listesi kullanılarak bir filtreleme yapıldı. Sonuç olarak 955 gen ismi kalmıştır.

**Hücre Hattı Özellik Vektörü Tablosu**

Yukarıda filtrelenen gen isimlerinin listeleri oluşturulup bu listelerle ilgili satırlar veri üzerinden elde edilmiştir.

Tüm tablolardaki indeks sütunu yeniden sıfır sayısından başlatılarak düzenlendi.

Sonuç tablosu 60 hücre hattı ve 3747 (897 + 963 + 955 + 932) gen ismi için boş değerli bir tablo oluşturuldu. Her bir hücre hattına ait değerler ilgili özellik verisinden çekilir ve tümü bir liste olarak birleştirilir. Bu 3747 uzunluğundaki bu liste sonuç tablosundaki ilgili hücre hattına ait satıra yerleştirilir. Eğer, herhangi bir özellik tipinden gelen veri içinde “NaN” etiketli değer bulunuyorsa bunun için farklı yöntemler izlenmiştir. Gen ifade verisi için, o gen ismine ait ortalama değer; mutasyon verisi için 0 değeri; metilasyon için ortalama değer atanmıştır.

Sonuç tablosu, boşluk değeri içermediği kontrol edilip hücre hattı sütunu alfabetik sıralandıktan sonra bir dosyaya yazdırılıp kaydedilmiştir.

### 8.3.14. GDSC – NCI-60 Çapraz Alan Analizindeki Modeller İçin Gereken Veri Setlerinin Oluşturulması

GDSC ve NCI-60 için çapraz alan analizinde kullanılmak üzere oluşturulan 3747 vektör uzunluğunda olan hücre hattı özellik tabloları ile beraber farklı modelleme senaryoları oluşturulmuştur. Öncelikle, kullanılacak olan bu iki veri arasındaki hücre hatlarının ortaklıkları incelenmiştir. NCI-60 isimleri için önceki isim düzenleme formatı göz önüne alınmıış, “nci” kodlu olanlardan bu kod silinmiş ve "7860", “a549atcc” isimleri sırasıyla “786o”, “a549” ile değiştirilmiştir. Sonuç olarak iki platform arasında 53 tane hattın ortak olduğu görülmüştür. NCI-60 verisi üzerine de bu değişiklikler aktarılmıştır. İki platforma ait olan ilaç yanıtı verileri üzerinden de hücre hattı ve ilaç ortaklıkları incelenmiştir. Ortak olan 53 hücre hattı ve 97 ilaç ismi olduğu görülmüştür.

### Analizde Kullanılacak Olan İlaç Yanıtı Verilerinin Düzenlenmesi

GDSC'de olup NCI-60'da olmayan 935 hücre hattı için ilaç yanıtı verisi üzerinden bu hatlara ait satırları almak üzere filtreleme yapılarak 255811 satırlı tablo senaryo 1 için oluşturuldu ve sonuç bir dosyaya yazdırıldı. Aynı şekilde, GDSC'de bulunup NCI-60'da olmayan 307 ilaç için ilaç yanıtı verisi üzerinden bu ilaçlara ait satırları almak üzere filtreleme yapılarak 318661 satırlı tablo senaryo 2 için oluşturuldu ve sonuç bir dosyaya yazdırıldı.

**Tablo 8.9.** GDSC – NCI-60 arasında oluşturulan çapraz alan analizi için oluşturulan senaryolarda kullanılacak olan verilerin düzenlenme planları.

| | **GDSC** | | **NCI-60** | |
|---|---|---|---|---|
| **Senaryo** | **Hücre Hattı Özellik Vektörü Dosyası** | **İlaç Yanıtı Dosyası** | **Hücre Hattı Özellik Vektörü Dosyası** | **İlaç Yanıtı Dosyası** |
| **1** | Ortak olmayan hücre hattı | Ortak olmayan hücre hattı | Tümü | Tümü |
| **2** | Ortak olmayan ilaçlara ait hücre hatları | Ortak olmayan ilaç | Tümü | Tümü |
| **3** | Tümü | Tümü | Ortak hücre hattı | Ortak hücre hattı |
| **4** | Tümü | Tümü | Ortak olmayan hücre hattı | Ortak olmayan hücre hattı |
| **5** | Tümü | Tümü | Ortak ilaçlara ait hücre hatları | Ortak ilaç |
| **6** | Tümü | Tümü | Ortak olmayan ilaçlara ait hücre hatları | Ortak olmayan ilaç |
| **7** | Tümü | Tümü | Tümü | Tümü |

Senaryo 1 ve 2 için NCI-60'ın tüm ilaç yanıtı verisi kullanılmıştır. Senaryo 3, 4, 5, 6, 7 için GDSC'nin tüm ilaç yanıtı verisi kullanılmıştır.

Ortak hücre hatları, NCI-60 – GDSC farkı hücre hatları, Ortak ilaçlar, NCI-60 – GDSC farkı ilaç listeleri için ilaç yanıtı tablosundan sırasıyla yapılan filtrelemelerle 53693, 6985, 5586, 55092 satırlı tablolar oluşturulup ayrı ayrı dosyalara yine sırasıyla senaryo 3, 4, 5, 6 için kaydedilmiştir.

**Analizde Kullanılacak Olan Hücre Hattı Özellik Vektörü Verilerinin Düzenlenmesi**

GDSC – NCI-60 farkı hücre hattı ve senaryo 2'deki hücre hattı listesi kullanılarak GDSC hücre hattı özellik vektörü dosyasından sırasıyla yapılan filtrelemelerle 935, 988 satırlı tablolar oluşturulup ayrı ayrı dosyalara yine sırasıyla senaryo 1, 2 için kaydedilmiştir.

Senaryo 1 ve 2 için NCI-60'ın tüm hücre hattı vektör verisi kullanılmıştır. Senaryo 3, 4, 5, 6, 7 için GDSC'nin tüm hücre hattı vektör verisi kullanılmıştır.

NCI-60 için oluşturulan senaryo 3, 4, 5, 6 ilaç yanıtı tablolarındaki hücre hattı listeleri çıkarılarak NCI-60 hücre hattı özellik vektöründeki hücre hattı sütunu üzerinden filtreleme yapılarak sırasıyla 53, 7, 60, 60 satırlı tablolar elde edilmiş ve senaryo 3 ve 4 için sonuçlar dosyalara yazdırılıp kaydedilmiştir. Senaryo 5 ve 6 için ise yine tablodaki tüm hücre hatları elde edildiğinden farklı bir dosyaya yazdırmaya gerek görülmemiştir.

### 8.3.15. GDSC - CCLE Arası Çapraz Alan Analizi Modellemeleri

Senaryo 1 için uygulanan yöntem diğer senaryolarda da benzer şekilde kullanılarak sonuçlar elde edilmiştir. Buna bağlı olarak Senaryo 1 haricindeki senaryoların yazımı için sadece uygulanan yöntemden farklı olan kısımlar belirtilmiştir.

**Senaryo 1 Modelleme Aşamaları**

Senaryo 1 için GDSC ve CCLE özelinde oluşturulan ilaç yanıtı ve hücre hattı özellik vektörlerine ait dosyalar modelde kullanılacak olan verinin oluşturulması için kullanıldı.

Öncelikle, tablolardaki hücre hattı listeleri kontrol edilmiş ve CCLE ilaç yanıtı verisi 503 hücre hattı için tekrar filtrelenmiştir. İki ilaç yanıtı verisi de hücre hattı ve ilaç ismi sütunlarına göre yeniden sıralanmıştır. İlaç isimleri ve parmak izleri listelere alınıp kontrol edilmiştir.

**Tablo 8.10.** GDSC - CCLE ÇAA Senaryo 1 için kullanılan verilerin genel özellikleri.

| | **Kullanılan veri** | **Tablo boyutu** | **Hücre hattı sayısı** | **İlaç sayısı** | **Parmak izi sayısı** |
|---|---|---|---|---|---|
| İlaç yanıtı verisi | GDSC_drug_response_df_for_cell_line_based_split_cross_domain_mode_1 | 327231, 6 | 612 | 404 | 203934 |
| İlaç yanıtı verisi | CCLE_drug_response_df_for_cross_domain | 11670, 6 | 504 | 24 | 11646 |
| Hücre hattı vektör verisi | GDSC_w_CCLE_cell_line_feature_vector_3747_df_for_cross_domain_mode_1 | 612, 3748 | 612 | - | - |
| Hücre hattı vektör verisi | CCLE_extracted_503_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform | 503, 3748 | 503 | - | - |

1024 rakamlı bir sayı olarak bulunan ilaç parmak izleri birer döngü yardımıyla 1024 elemanlı listelere çevrilmiştir. Bu yeni oluşturulan parmak izi listeleri GDSC ve CCLE için her bir satıra bir parmak izi olacak şekilde ayrı ayrı tablolar oluşturulmuştur. Oluşturulan tabloların ilk sütunlarına hücre hattı isimleri ve son sütunlarına pIC50 değerleri yerleştirildi. Hücre hattı özellik vektörü ve parmak izi tablosu birleştirileceğinden parmak izi tablosunun sütun isimleri ilgili vektör uzunluğu göz önüne alınıp yeni değerlerle değiştirilmiştir. Son adımda ise GDSC hücre hattı özellik vektörü ve

parmak izi tabloları hücre hattı isimleri üzerinden birleştirilmiştir ve CCLE için de aynı yöntem uygulanmıştır.

Modelleme kısmında, eğitim verisi GDSC için oluşturulan birleştirilmiş tablodan ilaç ismi, hücre hattı ismi ve pIC50 sütunları çıkartılarak oluşturuldu. Etiket verisi olarak ise aynı tablonun pIC50 sütunu kullanıldı. Test verisi, CCLE için oluşturulan birleştirilmiş tablodan ilaç ismi, hücre hattı ismi ve pIC50 sütunları çıkartılarak oluşturuldu. Test verisinin etiket verisi de pIC50 sütunundan alındı. Hem eğitim hem test verileri tek duyarlıklı kayan noktalı sayı biçimi (float32) olarak kaydedilmiş ve bu veri yapıları içinde boşluk değeri olup olmadığı kontrol edilmiştir.

Genel sonuç dosyası için parametreler, MAE, MSE, RMSE, nDGC (*normalized discounted cumulative gain*, normalize edilmiş indirimli kümülatif kazanç), Pearson korelasyonu, Spearman korelasyonu, $R^2$ metrikleri için sütun isimleri belirlenip bir dosya olarak kaydedilmiştir.

Python kütüphanelerinden os, joblib, sklearn (hem RandomForestRegressor hem metrikler için) rastgele orman modelini oluşturup kaydedebilmek için kullanılmıştır.

RandomForestRegressor metodu ile max_depth = 81, n_estimators = 100 (önceden uygulanan doku bazlı performans ölçümleri sonucu belirlenmiş olan en uygun değerler), verbose=3 (tüm çıktıların yazdırılması için uygulanan parametre değeri), n_jobs=6 (işlemci çekirdeği üzerinde kullanılan iş parçacığı sayısı), random_state=2 (modelin aynı şekilde kullanılabilmesi ve aynı sonuçları verebilmesi için kullanılan sabit sayı) parametreleri kullanılarak model özellikleri belirlenmiştir. Bu parametrelerle model, eğitim verisi ve ilgili etiket değerleri kullanılarak eğitilmiştir. Modelin tamamı joblib paketi yardımıyla bir dosyaya kaydedilmiştir.

Modelin CCLE test verisi üzerine tahmin sonuçlarını verebilmesi için predict metodu kullanılarak sonuçlar alınmıştır. Öncelikle, tahmin edilen sonuç değerleri CCLE gerçek değerleriyle yukarıda belirtilen metrikler için ayrı ayrı değerlendirilip parametre

değerleriyle beraber ilgili dosyaya ikinci satır olarak kaydedilmiştir. Sonraki adımda, hücre hattı ismi, ilaç ismi, gerçek pIC50, tahmin edilen pIC50 sütunları oluşturulup ilgili yerlere isim ve sayısal değerler eşleşen şekilde eklenmiş ve tablo bir dosyaya yazdırılarak kaydedilmiştir.

**Senaryo 2 Modelleme Aşamaları**

Senaryo 2 için GDSC ve CCLE özelinde oluşturulan ilaç yanıtı ve hücre hattı özellik vektörlerine ait dosyalar modelde kullanılacak olan verinin oluşturulması için kullanıldı. Öncelikle, tablolardaki hücre hattı listeleri kontrol edilmiş ve CCLE ilaç yanıtı verisi 503 hücre hattı için tekrar filtrelenmiştir. İki ilaç yanıtı verisi de hücre hattı ve ilaç ismi sütunlarına göre yeniden sıralanmıştır. İlaç isimleri ve parmak izleri listelere alınıp kontrol edilmiştir.

**Tablo 8.11.** GDSC - CCLE ÇAA Senaryo 2 için kullanılan verilerin genel özellikleri.

| | **Kullanılan veri** | **Tablo boyutu** | **Hücre hattı sayısı** | **İlaç sayısı** | **Parmak izi sayısı** |
|---|---|---|---|---|---|
| İlaç yanıtı verisi | GDSC_drug_response_df_for_drug_based_split_cross_domain_mode_2 | 327231, 6 | 612 | 391 | 203934 |
| İlaç yanıtı verisi | CCLE_drug_response_df_for_cross_domain | 11670, 6 | 504 | 24 | 11646 |
| Hücre hattı vektör verisi | GDSC_w_CCLE_cell_line_feature_vector_3747_df_for_cross_domain_mode_2 | 988, 3748 | 988 | - | - |
| Hücre hattı vektör verisi | CCLE_extracted_503_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform | 503, 3748 | 503 | - | - |

### Senaryo 3 Modelleme Aşamaları

Senaryo 3 için GDSC ve CCLE özelinde oluşturulan ilaç yanıtı ve hücre hattı özellik vektörlerine ait dosyalar modelde kullanılacak olan verinin oluşturulması için kullanıldı. Öncelikle, tablolardaki hücre hattı listeleri kontrol edilmiştir. İki ilaç yanıtı verisi de hücre hattı ve ilaç ismi sütunlarına göre yeniden sıralanmıştır. İlaç isimleri ve parmak izleri listelere alınıp kontrol edilmiştir.

**Tablo 8.12.** GDSC - CCLE ÇAA Senaryo 3 için kullanılan verilerin genel özellikleri.

| | **Kullanılan veri** | **Tablo boyutu** | **Hücre hattı sayısı** | **İlaç sayısı** | **Parmak izi sayısı** |
|---|---|---|---|---|---|
| İlaç yanıtı verisi | GDSC_drug_response_drugs_w_smiles_988cl_404dr_337K_v4_w_fps | 337761, 6 | 988 | 404 | 337761 |
| İlaç yanıtı verisi | CCLE_drug_response_df_for_cross_domain_mode_3 | 8666, 6 | 376 | 24 | 8666 |
| Hücre hattı vektör verisi | GDSC_extracted_988_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform | 988, 3748 | 988 | - | - |
| Hücre hattı vektör verisi | CCLE_w_GDSC_cell_line_3747_feature_vector_376_df_for_cross_domain_mode_3 | 376, 3748 | 376 | - | - |

### Senaryo 4 Modelleme Aşamaları

Senaryo 4 için GDSC ve CCLE özelinde oluşturulan ilaç yanıtı ve hücre hattı özellik vektörlerine ait dosyalar modelde kullanılacak olan verinin oluşturulması için kullanıldı. Öncelikle, tablolardaki hücre hattı listeleri kontrol edilmiştir ve CCLE ilaç yanıtı verisi 127

hücre hattı için tekrar filtrelenmiştir. İki ilaç yanıtı verisi de hücre hattı ve ilaç ismi sütunlarına göre yeniden sıralanmıştır. İlaç isimleri ve parmak izleri listelere alınıp kontrol edilmiştir.

**Tablo 8.13.** GDSC - CCLE ÇAA Senaryo 4 için kullanılan verilerin genel özellikleri.

| | **Kullanılan veri** | **Tablo boyutu** | **Hücre hattı sayısı** | **İlaç sayısı** | **Parmak izi sayısı** |
|---|---|---|---|---|---|
| İlaç yanıtı verisi | GDSC_drug_response_drugs_w_smiles_988cl_404dr_337K_v4_w_fps | 337761, 6 | 988 | 404 | 337761 |
| İlaç yanıtı verisi | CCLE_drug_response_df_for_cross_domain_mode_4 | 3004, 6 | 128 | 24 | 2980 |
| Hücre hattı vektör verisi | GDSC_extracted_988_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform | 988, 3748 | 988 | - | - |
| Hücre hattı vektör verisi | CCLE_w_GDSC_cell_line_3747_feature_vector_127_df_for_cross_domain_mode_4 | 376, 3748 | 127 | - | - |

**Senaryo 5 Modelleme Aşamaları**

Senaryo 5 için GDSC ve CCLE özelinde oluşturulan ilaç yanıtı ve hücre hattı özellik vektörlerine ait dosyalar modelde kullanılacak olan verinin oluşturulması için kullanıldı. Öncelikle, tablolardaki hücre hattı listeleri kontrol edilmiştir ve CCLE ilaç yanıtı verisi 503 hücre hattı için tekrar filtrelenmiştir. İki ilaç yanıtı verisi de hücre hattı ve ilaç ismi sütunlarına göre yeniden sıralanmıştır. İlaç isimleri ve parmak izleri listelere alınıp kontrol edilmiştir.

**Tablo 8.14.** GDSC - CCLE ÇAA Senaryo 5 için kullanılan verilerin genel özellikleri.

| | **Kullanılan veri** | **Tablo boyutu** | **Hücre hattı sayısı** | **İlaç sayısı** | **Parmak izi sayısı** |
|---|---|---|---|---|---|
| İlaç yanıtı verisi | GDSC_drug_response_drugs_w_smiles_988cl_404dr_337K_v4_w_fps | 337761, 6 | 988 | 404 | 337761 |
| İlaç yanıtı verisi | CCLE_drug_response_df_for_cross_domain_mode_5_v2 | 6265, 6 | 504 | 13 | 5246 |
| Hücre hattı vektör verisi | GDSC_extracted_988_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform | 988, 3748 | 988 | - | - |
| Hücre hattı vektör verisi | CCLE_extracted_503_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform | 503, 3748 | 503 | - | - |

**Senaryo 6 Modelleme Aşamaları**

Senaryo 6 için GDSC ve CCLE özelinde oluşturulan ilaç yanıtı ve hücre hattı özellik vektörlerine ait dosyalar modelde kullanılacak olan verinin oluşturulması için kullanıldı. Öncelikle, tablolardaki hücre hattı listeleri kontrol edilmiştir ve CCLE ilaç yanıtı verisi 503 hücre hattı için tekrar filtrelenmiştir. İki ilaç yanıtı verisi de hücre hattı ve ilaç ismi sütunlarına göre yeniden sıralanmıştır. İlaç isimleri ve parmak izleri listelere alınıp kontrol edilmiştir.

**Tablo 8.15.** GDSC - CCLE ÇAA Senaryo 6 için kullanılan verilerin genel özellikleri.

| | **Kullanılan veri** | **Tablo boyutu** | **Hücre hattı sayısı** | **İlaç sayısı** | **Parmak izi sayısı** |
|---|---|---|---|---|---|
| İlaç yanıtı verisi | GDSC_drug_response_drugs_w_smiles_988cl_404dr_337K_v4_w_fps | 337761, 6 | 988 | 404 | 337761 |
| İlaç yanıtı verisi | CCLE_drug_response_df_for_cross_domain_mode_6_v2 | 5405, 6 | 504 | 11 | 5394 |
| Hücre hattı vektör verisi | GDSC_extracted_988_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform | 988, 3748 | 988 | - | - |
| Hücre hattı vektör verisi | CCLE_extracted_503_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform | 503, 3748 | 503 | - | - |

**Senaryo 7 Modelleme Aşamaları**

Senaryo 7 için GDSC ve CCLE özelinde oluşturulan ilaç yanıtı ve hücre hattı özellik vektörlerine ait dosyalar modelde kullanılacak olan verinin oluşturulması için kullanıldı. Öncelikle, tablolardaki hücre hattı listeleri kontrol edilmiştir ve CCLE ilaç yanıtı verisi 503 hücre hattı için tekrar filtrelenmiştir. İki ilaç yanıtı verisi de hücre hattı ve ilaç ismi sütunlarına göre yeniden sıralanmıştır. İlaç isimleri ve parmak izleri listelere alınıp kontrol edilmiştir.

**Tablo 8.16.** GDSC - CCLE ÇAA Senaryo 7 için kullanılan verilerin genel özellikleri.

| | **Kullanılan veri** | **Tablo boyutu** | **Hücre hattı sayısı** | **İlaç sayısı** | **Parmak izi sayısı** |
|---|---|---|---|---|---|
| İlaç yanıtı verisi | GDSC_drug_response_drugs_w_smiles_988cl_404dr_337K_v4_w_fps | 337761, 6 | 988 | 404 | 337761 |
| İlaç yanıtı verisi | CCLE_drug_response_df_for_cross_domain | 11670, 6 | 504 | 24 | 11646 |
| Hücre hattı vektör verisi | GDSC_extracted_988_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform | 988, 3748 | 988 | - | - |
| Hücre hattı vektör verisi | CCLE_extracted_503_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform | 503, 3748 | 503 | - | - |

### 8.3.16. GDSC - NCI-60 Arası Çapraz Alan Analizi Modellemeleri

Senaryo 1 için uygulanan yöntem diğer senaryolarda da benzer şekilde kullanılarak sonuçlar elde edilmiştir. Buna bağlı olarak Senaryo 1 haricindeki senaryoların yazımı için sadece uygulanan yöntemden farklı olan kısımlar belirtilmiştir.

#### Senaryo 1 Modelleme Aşamaları

Senaryo 1 için GDSC ve NCI-60 özelinde oluşturulan ilaç yanıtı ve hücre hattı özellik vektörlerine ait dosyalar modelde kullanılacak olan verinin oluşturulması için kullanıldı. Öncelikle, tablolardaki hücre hattı listeleri kontrol edilmiştir. İki ilaç yanıtı verisi de hücre hattı ve ilaç ismi sütunlarına göre yeniden sıralanmıştır. İlaç isimleri ve parmak izleri listelere alınıp kontrol edilmiştir.

**Tablo 8.17.** GDSC - NCI-60 ÇAA Senaryo 1 için kullanılan verilerin genel özellikleri.

| | **Kullanılan veri** | **Tablo boyutu** | **Hücre hattı sayısı** | **İlaç sayısı** | **Parmak izi sayısı** |
|---|---|---|---|---|---|
| İlaç yanıtı verisi | GDSC_drug_response_df_for_cell_line_based_split_NCI-60_cross_domain_mode_1 | 318661, 6 | 935 | 404 | 318661 |
| İlaç yanıtı verisi | NCI-60_drug_response_rcellminer_drug_name_nsc_id_cl_name_pgi50_v3_edits_fps_drop | 60678, 7 | 60 | 1054 | 60678 |
| Hücre hattı vektör verisi | GDSC_w_NCI-60_cell_line_feature_vector_3747_df_for_cross_domain_mode_1 | 935, 3748 | 935 | - | - |
| Hücre hattı vektör verisi | NCI-60_cell_line_feature_vector_3747_df_for_cross_domain | 60, 3748 | 60 | - | - |

1024 rakamlı bir sayı olarak bulunan ilaç parmak izleri birer döngü yardımıyla 1024 elemanlı listelere çevrilmiştir. Bu yeni oluşturulan parmak izi listeleri GDSC ve NCI-60 için her bir satıra bir parmak izi olacak şekilde ayrı ayrı tablolar oluşturulmuştur. Oluşturulan tabloların ilk sütunlarına hücre hattı isimleri ve son sütunlarına pIC50 değerleri yerleştirildi. Hücre hattı özellik vektörü ve parmak izi tablosu birleştirileceğinden parmak izi tablosunun sütun isimleri ilgili vektör uzunluğu göz önüne alınıp yeni değerlerle değiştirilmiştir. Son adımda ise GDSC hücre hattı özellik vektörü ve parmak izi tabloları hücre hattı isimleri üzerinden birleştirilmiştir ve NCI-60 için de aynı yöntem uygulanmıştır.

Modelleme kısmında, eğitim verisi GDSC için oluşturulan birleştirilmiş tablodan ilaç ismi, hücre hattı ismi ve pIC50 sütunları çıkartılarak oluşturuldu. Etiket verisi olarak ise aynı tablonun pIC50 sütunu kullanıldı. Test verisi, NCI-60 için oluşturulan birleştirilmiş tablodan ilaç ismi, hücre hattı ismi ve pIC50 sütunları çıkartılarak oluşturuldu. Test verisinin etiket verisi de pIC50 sütunundan alındı. Hem eğitim hem test verileri tek

duyarlıklı kayan noktalı sayı biçimi (float32) olarak kaydedilmiş ve bu veri yapıları içinde boşluk değeri olup olmadığı kontrol edilmiştir.

Genel sonuç dosyası için parametreler, MAE, MSE, RMSE, nDGC, Pearson korelasyonu, Spearman korelasyonu, $R^2$ metrikleri için sütun isimleri belirlenip bir dosya olarak kaydedilmiştir.

Python kütüphanelerinden os, joblib, sklearn (hem RandomForestRegressor hem metrikler için) rastgele orman modelini oluşturup kaydedebilmek için kullanılmıştır.

RandomForestRegressor metodu ile max_depth = 81, n_estimators = 100 (önceden uygulanan doku bazlı performans ölçümleri sonucu belirlenmiş olan en uygun değerler), verbose=3 (tüm çıktıların yazdırılması için uygulanan parametre değeri), n_jobs=6 (işlemci çekirdeği üzerinde kullanılan iş parçacığı sayısı), random_state=2 (modelin aynı şekilde kullanılabilmesi ve aynı sonuçları verebilmesi için kullanılan sabit sayı) parametreleri kullanılarak model özellikleri belirlenmiştir. Bu parametrelerle model, eğitim verisi ve ilgili etiket değerleri kullanılarak eğitilmiştir. Modelin tamamı joblib paketi yardımıyla bir dosyaya kaydedilmiştir.

Modelin NCI-60 test verisi üzerine tahmin sonuçlarını verebilmesi için predict metodu kullanılarak sonuçlar alınmıştır. Öncelikle, tahmin edilen sonuç değerleri NCI-60 gerçek değerleriyle yukarıda belirtilen metrikler için ayrı ayrı değerlendirilip parametre değerleriyle beraber ilgili dosyaya ikinci satır olarak kaydedilmiştir. Sonraki adımda, hücre hattı ismi, ilaç ismi, gerçek pIC50, tahmin edilen pIC50 sütunları oluşturulup ilgili yerlere isim ve sayısal değerler eşleşen şekilde eklenmiş ve tablo bir dosyaya yazdırılarak kaydedilmiştir.

**Senaryo 2 Modelleme Aşamaları**

Senaryo 2 için GDSC ve NCI-60 özelinde oluşturulan ilaç yanıtı ve hücre hattı özellik vektörlerine ait dosyalar modelde kullanılacak olan verinin oluşturulması için kullanıldı. Öncelikle, tablolardaki hücre hattı listeleri kontrol edilmiştir. İki ilaç yanıtı verisi

de hücre hattı ve ilaç ismi sütunlarına göre yeniden sıralanmıştır. İlaç isimleri ve parmak izleri listelere alınıp kontrol edilmiştir.

**Tablo 8.18.** GDSC - NCI-60 ÇAA Senaryo 2 için kullanılan verilerin genel özellikleri.

| | **Kullanılan veri** | **Tablo boyutu** | **Hücre hattı sayısı** | **İlaç sayısı** | **Parmak izi sayısı** |
|---|---|---|---|---|---|
| İlaç yanıtı verisi | GDSC_drug_response_df_for_drug_based_split_cross_domain_mode_2 | 255811, 6 | 988 | 307 | 255811 |
| İlaç yanıtı verisi | NCI-60_drug_response_rcellminer_drug_name_nsc_id_cl_name_pgi50_v3_edits_fps_drop | 60768, 7 | 60 | 1054 | 60768 |
| Hücre hattı vektör verisi | GDSC_w_NCI-60_cell_line_feature_vector_3747_df_for_cross_domain_mode_2 | 988, 3748 | 988 | - | - |
| Hücre hattı vektör verisi | NCI-60_cell_line_feature_vector_3747_df_for_cross_domain | 60, 3748 | 60 | - | - |

**Senaryo 3 Modelleme Aşamaları**

Senaryo 3 için GDSC ve NCI-60 özelinde oluşturulan ilaç yanıtı ve hücre hattı özellik vektörlerine ait dosyalar modelde kullanılacak olan verinin oluşturulması için kullanıldı. Öncelikle, tablolardaki hücre hattı listeleri kontrol edilmiştir. İki ilaç yanıtı verisi de hücre hattı ve ilaç ismi sütunlarına göre yeniden sıralanmıştır. İlaç isimleri ve parmak izleri listelere alınıp kontrol edilmiştir.

**Tablo 8.19.** GDSC - NCI-60 ÇAA Senaryo 3 için kullanılan verilerin genel özellikleri.

| | **Kullanılan veri** | **Tablo boyutu** | **Hücre hattı sayısı** | **İlaç sayısı** | **Parmak izi sayısı** |
|---|---|---|---|---|---|
| İlaç yanıtı verisi | GDSC_drug_response_drugs_w_smiles_988cl_404dr_337K_v4_w_fps | 337761, 6 | 988 | 404 | 337761 |
| İlaç yanıtı verisi | NCI-60_drug_response_df_for_cross_domain_mode_3 | 53693, 7 | 53 | 1054 | 53693 |
| Hücre hattı vektör verisi | GDSC_extracted_988_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform | 988, 3748 | 988 | - | - |
| Hücre hattı vektör verisi | NCI-60_w_GDSC_cell_line_3747_feature_vector_53_df_for_cross_domain_mode_3 | 53, 3748 | 53 | - | - |

**Senaryo 4 Modelleme Aşamaları**

Senaryo 4 için GDSC ve NCI-60 özelinde oluşturulan ilaç yanıtı ve hücre hattı özellik vektörlerine ait dosyalar modelde kullanılacak olan verinin oluşturulması için kullanıldı. Öncelikle, tablolardaki hücre hattı listeleri kontrol edilmiştir. İki ilaç yanıtı verisi de hücre hattı ve ilaç ismi sütunlarına göre yeniden sıralanmıştır. İlaç isimleri ve parmak izleri listelere alınıp kontrol edilmiştir.

**Tablo 8.20.** GDSC - NCI-60 ÇAA Senaryo 4 için kullanılan verilerin genel özellikleri.

| | Kullanılan veri | Tablo boyutu | Hücre hattı sayısı | İlaç sayısı | Parmak izi sayısı |
|---|---|---|---|---|---|
| İlaç yanıtı verisi | GDSC_drug_response_drugs_w_smiles_988cl_404dr_337K_v4_w_fps | 337761, 6 | 988 | 404 | 337761 |
| İlaç yanıtı verisi | NCI-60_drug_response_df_for_cross_domain_mode_4 | 6985, 7 | 7 | 1054 | 6985 |
| Hücre hattı vektör verisi | GDSC_extracted_988_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform | 988, 3748 | 988 | - | - |
| Hücre hattı vektör verisi | NCI-60_w_GDSC_cell_line_3747_feature_vector_7_df_for_cross_domain_mode_4 | 7, 3748 | 7 | - | - |

### Senaryo 5 Modelleme Aşamaları

Senaryo 5 için GDSC ve NCI-60 özelinde oluşturulan ilaç yanıtı ve hücre hattı özellik vektörlerine ait dosyalar modelde kullanılacak olan verinin oluşturulması için kullanıldı. Öncelikle, tablolardaki hücre hattı listeleri kontrol edilmiştir. İki ilaç yanıtı verisi de hücre hattı ve ilaç ismi sütunlarına göre yeniden sıralanmıştır. İlaç isimleri ve parmak izleri listelere alınıp kontrol edilmiştir.

**Tablo 8.21.** GDSC - NCI-60 ÇAA Senaryo 5 için kullanılan verilerin genel özellikleri.

| | Kullanılan veri | Tablo boyutu | Hücre hattı sayısı | İlaç sayısı | Parmak izi sayısı |
|---|---|---|---|---|---|
| İlaç yanıtı verisi | GDSC_drug_response_drugs_w_smiles_988cl_404dr_337K_v4_w_fps | 337761, 6 | 988 | 404 | 337761 |
| İlaç yanıtı verisi | NCI-60_drug_response_df_for_cross_domain_mode_5_v2 | 5586, 6 | 60 | 97 | 5586 |
| Hücre hattı vektör verisi | GDSC_extracted_988_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform | 988, 3748 | 988 | - | - |
| Hücre hattı vektör verisi | NCI-60_w_GDSC_cell_line_feature_vector_3747_df_for_cross_domain | 60, 3748 | 60 | - | - |

**Senaryo 6 Modelleme Aşamaları**

Senaryo 6 için GDSC ve NCI-60 özelinde oluşturulan ilaç yanıtı ve hücre hattı özellik vektörlerine ait dosyalar modelde kullanılacak olan verinin oluşturulması için kullanıldı. Öncelikle, tablolardaki hücre hattı listeleri kontrol edilmiştir. İki ilaç yanıtı verisi de hücre hattı ve ilaç ismi sütunlarına göre yeniden sıralanmıştır. İlaç isimleri ve parmak izleri listelere alınıp kontrol edilmiştir.

**Tablo 8.22.** GDSC - NCI-60 ÇAA Senaryo 6 için kullanılan verilerin genel özellikleri.

| | **Kullanılan veri** | **Tablo boyutu** | **Hücre hattı sayısı** | **İlaç sayısı** | **Parmak izi sayısı** |
|---|---|---|---|---|---|
| İlaç yanıtı verisi | GDSC_drug_response_drugs_w_smiles_988cl_404dr_337K_v4_w_fps | 337761, 6 | 988 | 404 | 337761 |
| İlaç yanıtı verisi | NCI-60_drug_response_df_for_cross_domain_mode_6 | 55092, 6 | 60 | 957 | 55092 |
| Hücre hattı vektör verisi | GDSC_extracted_988_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform | 988, 3748 | 988 | - | - |
| Hücre hattı vektör verisi | NCI-60_w_GDSC_cell_line_feature_vector_3747_df_for_cross_domain | 60, 3748 | 60 | - | - |

**Senaryo 7 Modelleme Aşamaları**

Senaryo 7 için GDSC ve NCI-60 özelinde oluşturulan ilaç yanıtı ve hücre hattı özellik vektörlerine ait dosyalar modelde kullanılacak olan verinin oluşturulması için kullanıldı. Öncelikle, tablolardaki hücre hattı listeleri kontrol edilmiştir. İki ilaç yanıtı verisi de hücre hattı ve ilaç ismi sütunlarına göre yeniden sıralanmıştır. İlaç isimleri ve parmak izleri listelere alınıp kontrol edilmiştir.

**Tablo 8.23.** GDSC - NCI-60 ÇAA Senaryo 7 için kullanılan verilerin genel özellikleri.

| | **Kullanılan veri** | **Tablo boyutu** | **Hücre hattı sayısı** | **İlaç sayısı** | **Parmak izi sayısı** |
|---|---|---|---|---|---|
| İlaç yanıtı verisi | GDSC_drug_response_drugs_w_smiles_988cl_404dr_337K_v4_w_fps | 337761, 6 | 988 | 404 | 337761 |
| İlaç yanıtı verisi | NCI-60_drug_response_rcellminer_drug_name_nsc_id_cl_name_pgi50_v3_edits_fps_drop | 60678, 6 | 60 | 1054 | 60678 |
| Hücre hattı vektör verisi | GDSC_extracted_988_cell_lines_L1000_common_genes_3747_feature_vector_v1_selected_common_3_platform | 988, 3748 | 988 | - | - |
| Hücre hattı vektör verisi | NCI-60_w_GDSC_cell_line_feature_vector_3747_df_for_cross_domain | 60, 3748 | 60 | - | - |

## 8.4. EK-4

### 8.4.1. GDSC Verisi Kullanılarak Seçili Hücre Hatları Üzerinden Deepresponse-RF İle Tahmin Oluşturulması

DeepResponse-RF'nin GDSC sindirim sistemi dokusuna ait veriyle eğitilip KanSiL Lab envanterinde bulunan hücre hattı ve ilaçlardan oluşan çiftler için tahminler oluşturulmuştur. 420 çift için oluşturulan tahminleri içeren tablonun isim içerikli sütunlara göre alfabetik sıralanmış hali Tablo 8.24.'te gösterilmiştir.

**Tablo 8.24.** DeepResponse-RF ile GDSC sindirim sistemi verisi ile eğitilip KanSiL Lab hücre hattı – ilaç çiftleri için verilen ilaç yanıtı tahminleri.

| Hücre hattı Adı | İlaç Adı | DeepResponse-RF Tahmini (pIC50) | Gerçek Değer (pIC50) | Tahmini ve Gerçek Değer Farkı |
|---|---|---|---|---|
| CAMA-1 | Camptothecin | 6.719 | 6.274 | 0.445 |
| CAMA-1 | Cisplatin | 4.239 | 4.385 | 0.146 |
| CAMA-1 | Dactolisib | 6.443 | 7.151 | 0.708 |
| CAMA-1 | Fludarabine | 3.608 | 4.089 | 0.481 |
| CAMA-1 | PI-103 | 5.210 | 5.789 | 0.579 |
| CAMA-1 | Ruxolitinib | 4.192 | 4.048 | 0.144 |
| CAMA-1 | Selisistat | 3.630 | 3.826 | 0.196 |
| CAMA-1 | Sorafenib | 4.932 | 5.191 | 0.259 |
| CAMA-1 | Staurosporine | 7.176 | 6.514 | 0.662 |
| HCT-116 | Camptothecin | 7.209 | 7.297 | 0.087 |
| HCT-116 | Cisplatin | 5.001 | 5.222 | 0.221 |
| HCT-116 | Dactolisib | 6.889 | 6.813 | 0.076 |
| HCT-116 | Fludarabine | 3.979 | 3.952 | 0.027 |
| HCT-116 | PI-103 | 5.912 | 6.232 | 0.320 |
| HCT-116 | Ruxolitinib | 4.385 | 4.355 | 0.030 |
| HCT-116 | Selisistat | 3.875 | 3.846 | 0.029 |
| HCT-116 | Sorafenib | 5.090 | 5.093 | 0.004 |
| HCT-116 | Staurosporine | 7.504 | 7.582 | 0.079 |
| Hep3B2-1-7 | Camptothecin | 6.777 | 6.681 | 0.096 |
| Hep3B2-1-7 | Cisplatin | 4.765 | 4.876 | 0.111 |
| Hep3B2-1-7 | Dactolisib | 7.037 | 7.222 | 0.185 |
| Hep3B2-1-7 | Fludarabine | 4.098 | 4.138 | 0.039 |
| Hep3B2-1-7 | PI-103 | 4.913 | 4.754 | 0.159 |
| Hep3B2-1-7 | Ruxolitinib | 4.351 | 4.310 | 0.041 |

| | | | | |
|---|---|---|---|---|
| Hep3B2-1-7 | Selisistat | 3.955 | 3.990 | 0.035 |
| HuH-7 | Camptothecin | 5.536 | 5.234 | 0.302 |
| HuH-7 | Cisplatin | 3.764 | 3.697 | 0.066 |
| HuH-7 | Dactolisib | 6.051 | 6.062 | 0.011 |
| HuH-7 | Fludarabine | 3.218 | 3.027 | 0.191 |
| HuH-7 | PI-103 | 4.374 | 4.085 | 0.290 |
| HuH-7 | Ruxolitinib | 4.158 | 4.191 | 0.033 |
| HuH-7 | Selisistat | 3.674 | 3.764 | 0.090 |
| HuH-7 | Sorafenib | 5.081 | 5.371 | 0.290 |
| HuH-7 | Staurosporine | 6.380 | 5.963 | 0.417 |
| MCF7 | Camptothecin | 6.755 | 6.489 | 0.266 |
| MCF7 | Cisplatin | 4.253 | 4.009 | 0.244 |
| MCF7 | Dactolisib | 6.566 | 5.788 | 0.779 |
| MCF7 | Fludarabine | 3.993 | 3.255 | 0.738 |
| MCF7 | PI-103 | 5.066 | 3.872 | 1.194 |
| MCF7 | Ruxolitinib | 4.243 | 4.367 | 0.124 |
| MCF7 | Selisistat | 3.673 | 3.715 | 0.042 |
| MCF7 | Sorafenib | 4.821 | 5.204 | 0.383 |
| MCF7 | Staurosporine | 7.390 | 6.722 | 0.668 |
| MDA-MB-231 | Camptothecin | 6.762 | 6.967 | 0.205 |
| MDA-MB-231 | Cisplatin | 4.260 | 4.441 | 0.182 |
| MDA-MB-231 | Dactolisib | 6.646 | 6.902 | 0.255 |
| MDA-MB-231 | Fludarabine | 3.809 | 6.262 | 2.454 |
| MDA-MB-231 | PI-103 | 5.117 | 4.969 | 0.148 |
| MDA-MB-231 | Ruxolitinib | 4.299 | 4.127 | 0.172 |
| MDA-MB-231 | Selisistat | 3.699 | 3.781 | 0.082 |
| MDA-MB-231 | Sorafenib | 4.812 | 5.161 | 0.350 |
| MDA-MB-231 | Staurosporine | 7.590 | 8.016 | 0.427 |
| SK-HEP-1 | Camptothecin | 6.859 | 6.896 | 0.037 |
| SK-HEP-1 | Cisplatin | 4.346 | 4.299 | 0.047 |
| SK-HEP-1 | Dactolisib | 6.948 | 7.091 | 0.142 |
| SK-HEP-1 | Fludarabine | 3.998 | 4.037 | 0.039 |
| SK-HEP-1 | PI-103 | 4.429 | 4.254 | 0.175 |
| SK-HEP-1 | Ruxolitinib | 4.256 | 4.288 | 0.032 |
| SK-HEP-1 | Selisistat | 3.894 | 3.984 | 0.090 |
| SK-HEP-1 | Sorafenib | 4.961 | 5.044 | 0.083 |
| SK-HEP-1 | Staurosporine | 7.120 | 6.913 | 0.207 |
| SNU-182 | Camptothecin | 6.710 | 6.794 | 0.084 |
| SNU-182 | Cisplatin | 3.714 | 3.562 | 0.152 |
| SNU-182 | Dactolisib | 6.265 | 6.004 | 0.261 |
| SNU-182 | Fludarabine | 3.223 | 3.050 | 0.173 |

| | | | | |
|---|---|---|---|---|
| SNU-182 | PI-103 | 5.828 | 6.309 | 0.480 |
| SNU-182 | Ruxolitinib | 4.961 | 5.389 | 0.427 |
| SNU-182 | Selisistat | 3.617 | 3.584 | 0.033 |
| SNU-182 | Sorafenib | 4.516 | 4.397 | 0.119 |
| SNU-182 | Staurosporine | 7.727 | 7.824 | 0.097 |
| SNU-387 | Camptothecin | 5.667 | 5.249 | 0.418 |
| SNU-387 | Cisplatin | 3.621 | 3.320 | 0.302 |
| SNU-387 | Dactolisib | 6.435 | 6.509 | 0.074 |
| SNU-387 | Fludarabine | 3.424 | 3.378 | 0.046 |
| SNU-387 | PI-103 | 4.252 | 3.763 | 0.489 |
| SNU-387 | Ruxolitinib | 4.209 | 4.265 | 0.056 |
| SNU-387 | Selisistat | 3.785 | 3.835 | 0.050 |
| SNU-387 | Sorafenib | 4.162 | 3.587 | 0.575 |
| SNU-387 | Staurosporine | 7.585 | 7.638 | 0.053 |
| SNU-423 | Camptothecin | 6.887 | 6.831 | 0.056 |
| SNU-423 | Cisplatin | 4.569 | 4.522 | 0.047 |
| SNU-423 | Dactolisib | 6.832 | 6.766 | 0.066 |
| SNU-423 | Fludarabine | 4.062 | 4.165 | 0.103 |
| SNU-423 | PI-103 | 6.467 | 7.059 | 0.592 |
| SNU-423 | Ruxolitinib | 4.261 | 4.222 | 0.039 |
| SNU-423 | Selisistat | 3.687 | 3.611 | 0.076 |
| SNU-423 | Sorafenib | 4.609 | 4.350 | 0.258 |
| SNU-423 | Staurosporine | 7.705 | 7.742 | 0.037 |
| SNU-449 | Camptothecin | 5.936 | 5.739 | 0.197 |
| SNU-449 | Cisplatin | 3.941 | 3.936 | 0.005 |
| SNU-449 | Dactolisib | 6.434 | 6.488 | 0.054 |
| SNU-449 | Fludarabine | 3.044 | 2.955 | 0.089 |
| SNU-449 | PI-103 | 6.080 | 6.608 | 0.528 |
| SNU-449 | Ruxolitinib | 4.224 | 4.224 | 0.001 |
| SNU-449 | Selisistat | 3.548 | 3.447 | 0.102 |
| SNU-449 | Sorafenib | 4.533 | 4.382 | 0.151 |
| SNU-449 | Staurosporine | 6.874 | 6.517 | 0.356 |
| SNU-475 | PI-103 | 4.928 | 4.892 | 0.036 |
| SNU-475 | Ruxolitinib | 4.416 | 4.436 | 0.020 |
| SNU-475 | Selisistat | 3.855 | 3.854 | 0.001 |
| CAMA-1 | AG 490 | 4.527 | | 4.527 |
| CAMA-1 | AKTİ-1/2 | 5.180 | | 5.180 |
| CAMA-1 | Alsterpaulone | 4.489 | | 4.489 |
| CAMA-1 | Aspirin | 4.449 | | 4.449 |
| CAMA-1 | Brigatinib | 5.635 | | 5.635 |

| | | | | |
|---|---|---|---|---|
| CAMA-1 | Chloroquine Phosphate | 4.621 | | 4.621 |
| CAMA-1 | DAPT | 4.267 | | 4.267 |
| CAMA-1 | Doxorobicine | 6.205 | | 6.205 |
| CAMA-1 | Flurbiprofen | 4.764 | | 4.764 |
| CAMA-1 | Gemcitabine hyrocloride | 6.578 | | 6.578 |
| CAMA-1 | Ibuprofen | 4.865 | | 4.865 |
| CAMA-1 | Imatinib Mesylate | 6.745 | | 6.745 |
| CAMA-1 | KN-93 | 4.359 | | 4.359 |
| CAMA-1 | Larotrectinib | 5.142 | | 5.142 |
| CAMA-1 | Lenvatinib | 5.779 | | 5.779 |
| CAMA-1 | LY 294002 | 4.769 | | 4.769 |
| CAMA-1 | Naproxen | 4.358 | | 4.358 |
| CAMA-1 | PI 3-Kbeta Inhibitor VI | 4.574 | | 4.574 |
| CAMA-1 | PI 3-Kα Inhibitor VIII | 4.287 | | 4.287 |
| CAMA-1 | Pranlukast | 4.386 | | 4.386 |
| CAMA-1 | Rapamycsin | 6.792 | | 6.792 |
| CAMA-1 | Regorafenib | 5.038 | | 5.038 |
| CAMA-1 | Reparixin | 4.751 | | 4.751 |
| CAMA-1 | Sunitinib | 4.851 | | 4.851 |
| CAMA-1 | Thalidomide | 4.240 | | 4.240 |
| CAMA-1 | ZSTSK474 | 5.648 | | 5.648 |
| HCT-116 | AG 490 | 4.684 | | 4.684 |
| HCT-116 | AKTİ-1/2 | 5.430 | | 5.430 |
| HCT-116 | Alsterpaulone | 4.782 | | 4.782 |
| HCT-116 | Aspirin | 4.550 | | 4.550 |
| HCT-116 | Brigatinib | 5.523 | | 5.523 |
| HCT-116 | Chloroquine Phosphate | 5.017 | | 5.017 |
| HCT-116 | DAPT | 4.718 | | 4.718 |
| HCT-116 | Doxorobicine | 6.436 | | 6.436 |
| HCT-116 | Flurbiprofen | 5.016 | | 5.016 |
| HCT-116 | Gemcitabine hyrocloride | 7.418 | | 7.418 |
| HCT-116 | Ibuprofen | 5.029 | | 5.029 |
| HCT-116 | Imatinib Mesylate | 7.196 | | 7.196 |
| HCT-116 | KN-93 | 4.708 | | 4.708 |
| HCT-116 | Larotrectinib | 5.347 | | 5.347 |
| HCT-116 | Lenvatinib | 6.018 | | 6.018 |
| HCT-116 | LY 294002 | 4.849 | | 4.849 |

| | | | | |
|---|---|---|---|---|
| HCT-116 | Naproxen | 4.644 | | 4.644 |
| HCT-116 | PI 3-Kbeta Inhibitor VI | 4.827 | | 4.827 |
| HCT-116 | PI 3-Kα Inhibitor VIII | 4.644 | | 4.644 |
| HCT-116 | Pranlukast | 4.940 | | 4.940 |
| HCT-116 | Rapamycsin | 6.821 | | 6.821 |
| HCT-116 | Regorafenib | 5.221 | | 5.221 |
| HCT-116 | Reparixin | 5.222 | | 5.222 |
| HCT-116 | Sunitinib | 5.230 | | 5.230 |
| HCT-116 | Thalidomide | 4.541 | | 4.541 |
| HCT-116 | ZSTSK474 | 5.502 | | 5.502 |
| Hep3B2-1-7 | AG 490 | 4.631 | | 4.631 |
| Hep3B2-1-7 | AKTİ-1/2 | 5.224 | | 5.224 |
| Hep3B2-1-7 | Alsterpaulone | 4.790 | | 4.790 |
| Hep3B2-1-7 | Aspirin | 4.431 | | 4.431 |
| Hep3B2-1-7 | Brigatinib | 5.695 | | 5.695 |
| Hep3B2-1-7 | Chloroquine Phosphate | 4.834 | | 4.834 |
| Hep3B2-1-7 | DAPT | 4.518 | | 4.518 |
| Hep3B2-1-7 | Doxorobicine | 6.195 | | 6.195 |
| Hep3B2-1-7 | Flurbiprofen | 4.954 | | 4.954 |
| Hep3B2-1-7 | Gemcitabine hyrocloride | 6.324 | | 6.324 |
| Hep3B2-1-7 | Ibuprofen | 4.874 | | 4.874 |
| Hep3B2-1-7 | Imatinib Mesylate | 6.772 | | 6.772 |
| Hep3B2-1-7 | KN-93 | 4.640 | | 4.640 |
| Hep3B2-1-7 | Larotrectinib | 5.433 | | 5.433 |
| Hep3B2-1-7 | Lenvatinib | 6.030 | | 6.030 |
| Hep3B2-1-7 | LY 294002 | 5.052 | | 5.052 |
| Hep3B2-1-7 | Naproxen | 4.575 | | 4.575 |
| Hep3B2-1-7 | PI 3-Kbeta Inhibitor VI | 4.693 | | 4.693 |
| Hep3B2-1-7 | PI 3-Kα Inhibitor VIII | 4.508 | | 4.508 |
| Hep3B2-1-7 | Pranlukast | 4.541 | | 4.541 |
| Hep3B2-1-7 | Rapamycsin | 6.781 | | 6.781 |
| Hep3B2-1-7 | Regorafenib | 5.154 | | 5.154 |
| Hep3B2-1-7 | Reparixin | 4.990 | | 4.990 |
| Hep3B2-1-7 | Sorafenib | 5.014 | | 5.014 |
| Hep3B2-1-7 | Staurosporine | 6.789 | | 6.789 |
| Hep3B2-1-7 | Sunitinib | 5.062 | | 5.062 |
| Hep3B2-1-7 | Thalidomide | 4.671 | | 4.671 |
| Hep3B2-1-7 | ZSTSK474 | 5.300 | | 5.300 |
| HuH-7 | AG 490 | 4.326 | | 4.326 |

| | | | | |
|---|---|---|---|---|
| HuH-7 | AKTİ-1/2 | 5.152 | | 5.152 |
| HuH-7 | Alsterpaulone | 4.440 | | 4.440 |
| HuH-7 | Aspirin | 4.006 | | 4.006 |
| HuH-7 | Brigatinib | 5.578 | | 5.578 |
| HuH-7 | Chloroquine Phosphate | 4.518 | | 4.518 |
| HuH-7 | DAPT | 4.109 | | 4.109 |
| HuH-7 | Doxorobicine | 5.584 | | 5.584 |
| HuH-7 | Flurbiprofen | 4.494 | | 4.494 |
| HuH-7 | Gemcitabine hyrocloride | 6.400 | | 6.400 |
| HuH-7 | Ibuprofen | 4.593 | | 4.593 |
| HuH-7 | Imatinib Mesylate | 6.023 | | 6.023 |
| HuH-7 | KN-93 | 4.180 | | 4.180 |
| HuH-7 | Larotrectinib | 4.938 | | 4.938 |
| HuH-7 | Lenvatinib | 5.730 | | 5.730 |
| HuH-7 | LY 294002 | 4.925 | | 4.925 |
| HuH-7 | Naproxen | 4.209 | | 4.209 |
| HuH-7 | PI 3-Kbeta Inhibitor VI | 4.653 | | 4.653 |
| HuH-7 | PI 3-Kα Inhibitor VIII | 4.212 | | 4.212 |
| HuH-7 | Pranlukast | 4.202 | | 4.202 |
| HuH-7 | Rapamycsin | 6.697 | | 6.697 |
| HuH-7 | Regorafenib | 5.025 | | 5.025 |
| HuH-7 | Reparixin | 4.456 | | 4.456 |
| HuH-7 | Sunitinib | 4.566 | | 4.566 |
| HuH-7 | Thalidomide | 4.088 | | 4.088 |
| HuH-7 | ZSTSK474 | 5.295 | | 5.295 |
| MCF7 | AG 490 | 4.420 | | 4.420 |
| MCF7 | AKTİ-1/2 | 5.045 | | 5.045 |
| MCF7 | Alsterpaulone | 4.651 | | 4.651 |
| MCF7 | Aspirin | 4.314 | | 4.314 |
| MCF7 | Brigatinib | 5.617 | | 5.617 |
| MCF7 | Chloroquine Phosphate | 4.744 | | 4.744 |
| MCF7 | DAPT | 4.245 | | 4.245 |
| MCF7 | Doxorobicine | 5.978 | | 5.978 |
| MCF7 | Flurbiprofen | 4.535 | | 4.535 |
| MCF7 | Gemcitabine hyrocloride | 6.623 | | 6.623 |
| MCF7 | Ibuprofen | 4.681 | | 4.681 |
| MCF7 | Imatinib Mesylate | 6.799 | | 6.799 |

| | | | | |
|---|---|---|---|---|
| MCF7 | KN-93 | 4.457 | | 4.457 |
| MCF7 | Larotrectinib | 4.954 | | 4.954 |
| MCF7 | Lenvatinib | 5.641 | | 5.641 |
| MCF7 | LY 294002 | 4.751 | | 4.751 |
| MCF7 | Naproxen | 4.217 | | 4.217 |
| MCF7 | PI 3-Kbeta Inhibitor VI | 4.412 | | 4.412 |
| MCF7 | PI 3-Kα Inhibitor VIII | 4.447 | | 4.447 |
| MCF7 | Pranlukast | 4.467 | | 4.467 |
| MCF7 | Rapamycsin | 6.621 | | 6.621 |
| MCF7 | Regorafenib | 4.866 | | 4.866 |
| MCF7 | Reparixin | 4.716 | | 4.716 |
| MCF7 | Sunitinib | 4.741 | | 4.741 |
| MCF7 | Thalidomide | 4.429 | | 4.429 |
| MCF7 | ZSTSK474 | 5.442 | | 5.442 |
| MDA-MB-231 | AG 490 | 4.455 | | 4.455 |
| MDA-MB-231 | AKTİ-1/2 | 5.137 | | 5.137 |
| MDA-MB-231 | Alsterpaulone | 4.595 | | 4.595 |
| MDA-MB-231 | Aspirin | 4.354 | | 4.354 |
| MDA-MB-231 | Brigatinib | 5.501 | | 5.501 |
| MDA-MB-231 | Chloroquine Phosphate | 4.611 | | 4.611 |
| MDA-MB-231 | DAPT | 4.328 | | 4.328 |
| MDA-MB-231 | Doxorobicine | 6.071 | | 6.071 |
| MDA-MB-231 | Flurbiprofen | 4.649 | | 4.649 |
| MDA-MB-231 | Gemcitabine hyrocloride | 6.774 | | 6.774 |
| MDA-MB-231 | Ibuprofen | 4.652 | | 4.652 |
| MDA-MB-231 | Imatinib Mesylate | 6.497 | | 6.497 |
| MDA-MB-231 | KN-93 | 4.480 | | 4.480 |
| MDA-MB-231 | Larotrectinib | 4.984 | | 4.984 |
| MDA-MB-231 | Lenvatinib | 5.713 | | 5.713 |
| MDA-MB-231 | LY 294002 | 4.773 | | 4.773 |
| MDA-MB-231 | Naproxen | 4.232 | | 4.232 |
| MDA-MB-231 | PI 3-Kbeta Inhibitor VI | 4.489 | | 4.489 |
| MDA-MB-231 | PI 3-Kα Inhibitor VIII | 4.472 | | 4.472 |
| MDA-MB-231 | Pranlukast | 4.519 | | 4.519 |
| MDA-MB-231 | Rapamycsin | 6.742 | | 6.742 |
| MDA-MB-231 | Regorafenib | 4.899 | | 4.899 |
| MDA-MB-231 | Reparixin | 4.719 | | 4.719 |
| MDA-MB-231 | Sunitinib | 4.919 | | 4.919 |
| MDA-MB-231 | Thalidomide | 4.340 | | 4.340 |

| | | | | |
|---|---|---|---|---|
| MDA-MB-231 | ZSTSK474 | 5.347 | | 5.347 |
| SK-HEP-1 | AG 490 | 4.697 | | 4.697 |
| SK-HEP-1 | AKTİ-1/2 | 4.902 | | 4.902 |
| SK-HEP-1 | Alsterpaulone | 4.512 | | 4.512 |
| SK-HEP-1 | Aspirin | 4.509 | | 4.509 |
| SK-HEP-1 | Brigatinib | 5.537 | | 5.537 |
| SK-HEP-1 | Chloroquine Phosphate | 4.703 | | 4.703 |
| SK-HEP-1 | DAPT | 4.294 | | 4.294 |
| SK-HEP-1 | Doxorobicine | 6.235 | | 6.235 |
| SK-HEP-1 | Flurbiprofen | 4.572 | | 4.572 |
| SK-HEP-1 | Gemcitabine hyrocloride | 7.586 | | 7.586 |
| SK-HEP-1 | Ibuprofen | 4.631 | | 4.631 |
| SK-HEP-1 | Imatinib Mesylate | 7.060 | | 7.060 |
| SK-HEP-1 | KN-93 | 4.488 | | 4.488 |
| SK-HEP-1 | Larotrectinib | 5.056 | | 5.056 |
| SK-HEP-1 | Lenvatinib | 5.856 | | 5.856 |
| SK-HEP-1 | LY 294002 | 4.595 | | 4.595 |
| SK-HEP-1 | Naproxen | 4.295 | | 4.295 |
| SK-HEP-1 | PI 3-Kbeta Inhibitor VI | 4.419 | | 4.419 |
| SK-HEP-1 | PI 3-Kα Inhibitor VIII | 4.389 | | 4.389 |
| SK-HEP-1 | Pranlukast | 4.471 | | 4.471 |
| SK-HEP-1 | Rapamycsin | 6.836 | | 6.836 |
| SK-HEP-1 | Regorafenib | 5.011 | | 5.011 |
| SK-HEP-1 | Reparixin | 4.783 | | 4.783 |
| SK-HEP-1 | Sunitinib | 4.773 | | 4.773 |
| SK-HEP-1 | Thalidomide | 4.343 | | 4.343 |
| SK-HEP-1 | ZSTSK474 | 4.578 | | 4.578 |
| SNU-182 | AG 490 | 4.442 | | 4.442 |
| SNU-182 | AKTİ-1/2 | 5.249 | | 5.249 |
| SNU-182 | Alsterpaulone | 4.457 | | 4.457 |
| SNU-182 | Aspirin | 4.164 | | 4.164 |
| SNU-182 | Brigatinib | 5.483 | | 5.483 |
| SNU-182 | Chloroquine Phosphate | 4.619 | | 4.619 |
| SNU-182 | DAPT | 4.161 | | 4.161 |
| SNU-182 | Doxorobicine | 6.141 | | 6.141 |
| SNU-182 | Flurbiprofen | 4.457 | | 4.457 |
| SNU-182 | Gemcitabine hyrocloride | 6.836 | | 6.836 |

| | | | | |
|---|---|---|---|---|
| SNU-182 | Ibuprofen | 4.438 | | 4.438 |
| SNU-182 | Imatinib Mesylate | 6.543 | | 6.543 |
| SNU-182 | KN-93 | 4.368 | | 4.368 |
| SNU-182 | Larotrectinib | 4.853 | | 4.853 |
| SNU-182 | Lenvatinib | 5.647 | | 5.647 |
| SNU-182 | LY 294002 | 4.709 | | 4.709 |
| SNU-182 | Naproxen | 4.205 | | 4.205 |
| SNU-182 | PI 3-Kbeta Inhibitor VI | 4.582 | | 4.582 |
| SNU-182 | PI 3-Kα Inhibitor VIII | 4.336 | | 4.336 |
| SNU-182 | Pranlukast | 4.501 | | 4.501 |
| SNU-182 | Rapamycsin | 6.366 | | 6.366 |
| SNU-182 | Regorafenib | 4.711 | | 4.711 |
| SNU-182 | Reparixin | 4.739 | | 4.739 |
| SNU-182 | Sunitinib | 4.793 | | 4.793 |
| SNU-182 | Thalidomide | 4.271 | | 4.271 |
| SNU-182 | ZSTSK474 | 5.840 | | 5.840 |
| SNU-387 | AG 490 | 4.547 | | 4.547 |
| SNU-387 | AKTİ-1/2 | 5.423 | | 5.423 |
| SNU-387 | Alsterpaulone | 4.368 | | 4.368 |
| SNU-387 | Aspirin | 4.123 | | 4.123 |
| SNU-387 | Brigatinib | 5.541 | | 5.541 |
| SNU-387 | Chloroquine Phosphate | 4.627 | | 4.627 |
| SNU-387 | DAPT | 4.083 | | 4.083 |
| SNU-387 | Doxorobicine | 5.892 | | 5.892 |
| SNU-387 | Flurbiprofen | 4.498 | | 4.498 |
| SNU-387 | Gemcitabine hyrocloride | 6.153 | | 6.153 |
| SNU-387 | Ibuprofen | 4.586 | | 4.586 |
| SNU-387 | Imatinib Mesylate | 6.684 | | 6.684 |
| SNU-387 | KN-93 | 4.344 | | 4.344 |
| SNU-387 | Larotrectinib | 4.969 | | 4.969 |
| SNU-387 | Lenvatinib | 5.589 | | 5.589 |
| SNU-387 | LY 294002 | 4.590 | | 4.590 |
| SNU-387 | Naproxen | 4.198 | | 4.198 |
| SNU-387 | PI 3-Kbeta Inhibitor VI | 4.472 | | 4.472 |
| SNU-387 | PI 3-Kα Inhibitor VIII | 4.301 | | 4.301 |
| SNU-387 | Pranlukast | 4.467 | | 4.467 |
| SNU-387 | Rapamycsin | 6.413 | | 6.413 |
| SNU-387 | Regorafenib | 4.489 | | 4.489 |
| SNU-387 | Reparixin | 4.695 | | 4.695 |

| | | | | |
|---|---|---|---|---|
| SNU-387 | Sunitinib | 4.730 | | 4.730 |
| SNU-387 | Thalidomide | 4.167 | | 4.167 |
| SNU-387 | ZSTSK474 | 5.970 | | 5.970 |
| SNU-423 | AG 490 | 4.711 | | 4.711 |
| SNU-423 | AKTİ-1/2 | 5.329 | | 5.329 |
| SNU-423 | Alsterpaulone | 4.598 | | 4.598 |
| SNU-423 | Aspirin | 4.574 | | 4.574 |
| SNU-423 | Brigatinib | 5.510 | | 5.510 |
| SNU-423 | Chloroquine Phosphate | 4.818 | | 4.818 |
| SNU-423 | DAPT | 4.578 | | 4.578 |
| SNU-423 | Doxorobicine | 6.201 | | 6.201 |
| SNU-423 | Flurbiprofen | 4.681 | | 4.681 |
| SNU-423 | Gemcitabine hyrocloride | 7.131 | | 7.131 |
| SNU-423 | Ibuprofen | 4.766 | | 4.766 |
| SNU-423 | Imatinib Mesylate | 7.019 | | 7.019 |
| SNU-423 | KN-93 | 4.716 | | 4.716 |
| SNU-423 | Larotrectinib | 5.145 | | 5.145 |
| SNU-423 | Lenvatinib | 5.753 | | 5.753 |
| SNU-423 | LY 294002 | 4.799 | | 4.799 |
| SNU-423 | Naproxen | 4.483 | | 4.483 |
| SNU-423 | PI 3-Kbeta Inhibitor VI | 4.855 | | 4.855 |
| SNU-423 | PI 3-Kα Inhibitor VIII | 4.676 | | 4.676 |
| SNU-423 | Pranlukast | 4.623 | | 4.623 |
| SNU-423 | Rapamycsin | 6.686 | | 6.686 |
| SNU-423 | Regorafenib | 4.798 | | 4.798 |
| SNU-423 | Reparixin | 4.917 | | 4.917 |
| SNU-423 | Sunitinib | 4.960 | | 4.960 |
| SNU-423 | Thalidomide | 4.354 | | 4.354 |
| SNU-423 | ZSTSK474 | 6.157 | | 6.157 |
| SNU-449 | AG 490 | 4.579 | | 4.579 |
| SNU-449 | AKTİ-1/2 | 4.825 | | 4.825 |
| SNU-449 | Alsterpaulone | 4.302 | | 4.302 |
| SNU-449 | Aspirin | 4.062 | | 4.062 |
| SNU-449 | Brigatinib | 5.526 | | 5.526 |
| SNU-449 | Chloroquine Phosphate | 4.578 | | 4.578 |
| SNU-449 | DAPT | 4.223 | | 4.223 |
| SNU-449 | Doxorobicine | 5.181 | | 5.181 |
| SNU-449 | Flurbiprofen | 4.527 | | 4.527 |

| | | | | |
|---|---|---|---|---|
| SNU-449 | Gemcitabine hyrocloride | 6.949 | | 6.949 |
| SNU-449 | Ibuprofen | 4.545 | | 4.545 |
| SNU-449 | Imatinib Mesylate | 6.883 | | 6.883 |
| SNU-449 | KN-93 | 4.271 | | 4.271 |
| SNU-449 | Larotrectinib | 4.847 | | 4.847 |
| SNU-449 | Lenvatinib | 5.685 | | 5.685 |
| SNU-449 | LY 294002 | 4.692 | | 4.692 |
| SNU-449 | Naproxen | 4.286 | | 4.286 |
| SNU-449 | PI 3-Kbeta Inhibitor VI | 4.632 | | 4.632 |
| SNU-449 | PI 3-Kα Inhibitor VIII | 4.262 | | 4.262 |
| SNU-449 | Pranlukast | 4.420 | | 4.420 |
| SNU-449 | Rapamycsin | 6.824 | | 6.824 |
| SNU-449 | Regorafenib | 4.728 | | 4.728 |
| SNU-449 | Reparixin | 4.652 | | 4.652 |
| SNU-449 | Sunitinib | 4.749 | | 4.749 |
| SNU-449 | Thalidomide | 4.126 | | 4.126 |
| SNU-449 | ZSTSK474 | 5.612 | | 5.612 |
| SNU-475 | AG 490 | 4.457 | | 4.457 |
| SNU-475 | AKTİ-1/2 | 4.930 | | 4.930 |
| SNU-475 | Alsterpaulone | 4.740 | | 4.740 |
| SNU-475 | Aspirin | 4.476 | | 4.476 |
| SNU-475 | Brigatinib | 5.554 | | 5.554 |
| SNU-475 | Camptothecin | 6.592 | | 6.592 |
| SNU-475 | Chloroquine Phosphate | 4.651 | | 4.651 |
| SNU-475 | Cisplatin | 4.519 | | 4.519 |
| SNU-475 | Dactolisib | 6.675 | | 6.675 |
| SNU-475 | DAPT | 4.533 | | 4.533 |
| SNU-475 | Doxorobicine | 6.083 | | 6.083 |
| SNU-475 | Fludarabine | 3.845 | | 3.845 |
| SNU-475 | Flurbiprofen | 4.646 | | 4.646 |
| SNU-475 | Gemcitabine hyrocloride | 6.134 | | 6.134 |
| SNU-475 | Ibuprofen | 4.637 | | 4.637 |
| SNU-475 | Imatinib Mesylate | 6.553 | | 6.553 |
| SNU-475 | KN-93 | 4.642 | | 4.642 |
| SNU-475 | Larotrectinib | 5.079 | | 5.079 |
| SNU-475 | Lenvatinib | 6.051 | | 6.051 |
| SNU-475 | LY 294002 | 4.924 | | 4.924 |
| SNU-475 | Naproxen | 4.352 | | 4.352 |

| | | | | |
|---|---|---|---|---|
| SNU-475 | PI 3-Kbeta Inhibitor VI | 4.593 | | 4.593 |
| SNU-475 | PI 3-Kα Inhibitor VIII | 4.539 | | 4.539 |
| SNU-475 | Pranlukast | 4.569 | | 4.569 |
| SNU-475 | Rapamycsin | 6.504 | | 6.504 |
| SNU-475 | Regorafenib | 5.058 | | 5.058 |
| SNU-475 | Reparixin | 4.867 | | 4.867 |
| SNU-475 | Sorafenib | 4.963 | | 4.963 |
| SNU-475 | Staurosporine | 7.510 | | 7.510 |
| SNU-475 | Sunitinib | 4.956 | | 4.956 |
| SNU-475 | Thalidomide | 4.621 | | 4.621 |
| SNU-475 | ZSTSK474 | 5.290 | | 5.290 |
| | **Ortalama:** | **5.072** | **5.111** | **3.909** |
| | | **DeepResponse-RF Tahmini (pIC50)** | **Gerçek Değer (pIC50)** | **Tahmini ve Gerçek Değer Farkı** |

### 8.4.2. GDSC Verisi Kullanılarak Eğitilen Deepresponse-RF İle DrugBank İlaçları İçin İlaç Yanıtı Tahmini Üretimi

GDSC sindirim sistemi verisiyle eğitilen DeepResponse-RF modeliyle, aynı veride bulunan hücre hatları ve DrugBank veri tabanına ait ilaçların kombinasyonları için ilaç yanıtı tahminleri oluşturulmuştur. Bu tahminlerin (en yüksek değerli 100 çift için) en yüksek ilaç yanıt değerine göre sıralanmış hali Tablo 8.25.'te verilmiştir.

**Tablo 8.25.** GDSC sindirim sistemi verisiyle ile eğitilmiş DeepResponse-RF modeli üzerinden tahmin verilen DrugBank veri tabanındaki ilaçlar.

| Hücre Hattı Adı | DrugBank İlaç Adı | DeepResponse-RF Tahmini (pIC50) |
|---|---|---|
| SNU-398 | Patupilone | 9.296232 |
| HGC-27 | Patupilone | 9.034492 |
| RKO | Patupilone | 9.028823 |
| HLE | Patupilone | 8.989822 |
| T84 | Patupilone | 8.953306 |
| HCT-15 | Patupilone | 8.882134 |
| SW620 | Patupilone | 8.873312 |
| JHH-7 | Patupilone | 8.851687 |
| HuTu-80 | Patupilone | 8.833986 |
| HuTu-80 | Elesclomol | 8.749374 |
| HCC2998 | Patupilone | 8.748809 |
| HT-29 | Patupilone | 8.736153 |
| MKN28 | Patupilone | 8.730866 |

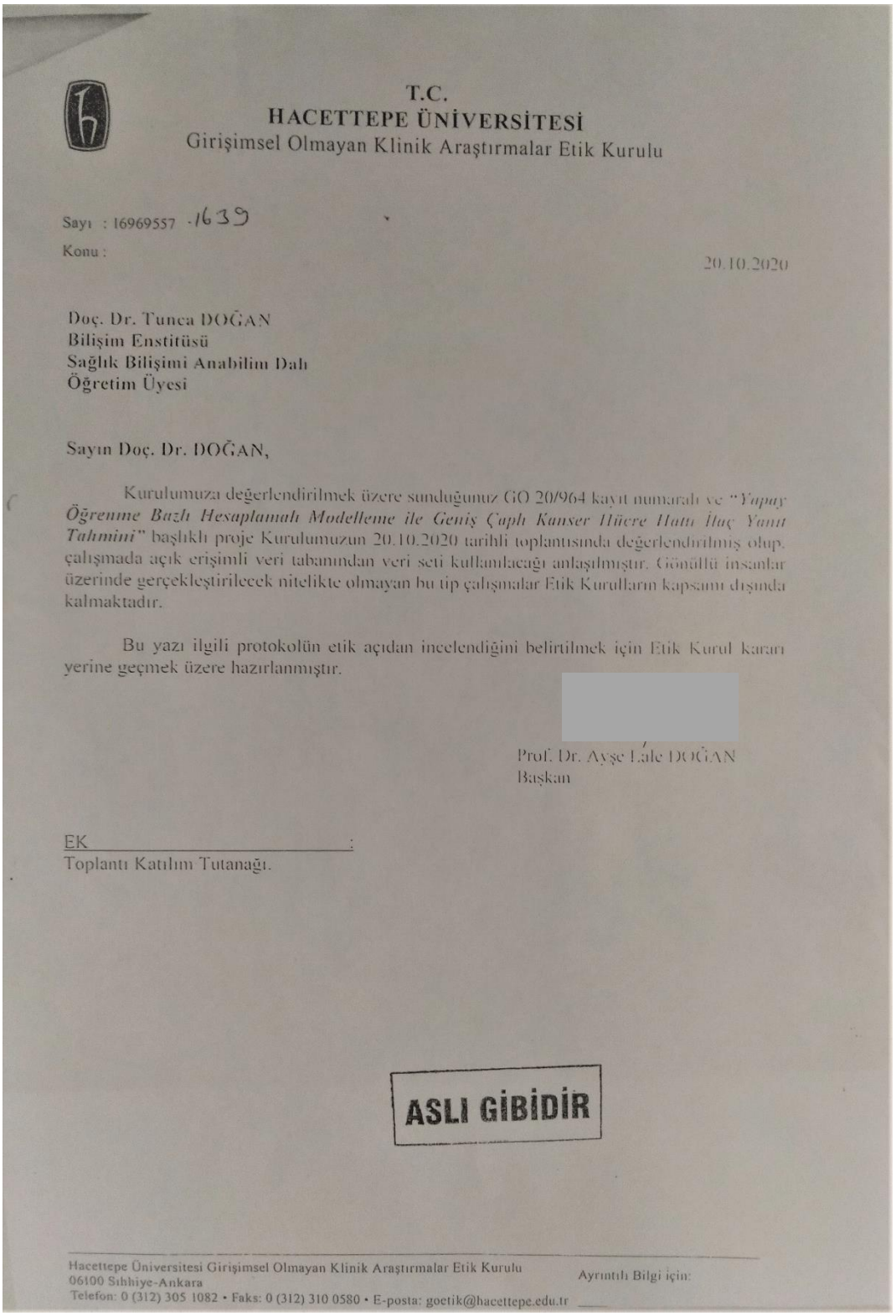
| | | |
|---|---|---|
| ETK-1 | Patupilone | 8.729611 |
| SNU-398 | Mipsagargin | 8.683044 |
| NUGC-3 | Patupilone | 8.676121 |
| SNU-423 | Patupilone | 8.672131 |
| SK-HEP-1 | Patupilone | 8.665949 |
| COLO-205 | Patupilone | 8.657949 |
| LS-180 | Patupilone | 8.646618 |
| HGC-27 | Daporinad | 8.637592 |
| HGC-27 | Mipsagargin | 8.60635 |
| LS-1034 | Patupilone | 8.591354 |
| LoVo | Patupilone | 8.576564 |
| AGS | Patupilone | 8.531659 |
| SNU-449 | Patupilone | 8.528723 |
| JHH-6 | Patupilone | 8.521119 |
| NCI-H747 | Patupilone | 8.488531 |
| Hep3B2-1-7 | Patupilone | 8.476638 |
| HSC-39 | Daporinad | 8.467811 |
| SNU-398 | Ixabepilone | 8.467151 |
| CCK-81 | Patupilone | 8.46628 |
| COLO-320-HSR | Daporinad | 8.452111 |
| HCT-116 | AZD-4877 | 8.446659 |
| SNU-175 | Patupilone | 8.437978 |
| MKN45 | Patupilone | 8.427518 |
| LS-123 | Patupilone | 8.423864 |
| KM12 | Patupilone | 8.420902 |
| HSC-39 | Docetaxel | 8.419474 |
| SNU-175 | Vinblastine | 8.417462 |
| IM-95 | Patupilone | 8.408554 |
| C3A | Patupilone | 8.408402 |
| HGC-27 | Bortezomib | 8.408206 |
| HuCCT1 | Patupilone | 8.39367 |
| HGC-27 | Ixabepilone | 8.392745 |
| HCT-116 | Filanesib | 8.391697 |
| AGS | AZD-4877 | 8.382487 |
| LS-411N | Patupilone | 8.38051 |
| JHH-4 | Patupilone | 8.379844 |
| COLO-320-HSR | Patupilone | 8.376911 |
| SNU-175 | Vindesine | 8.372679 |
| HT-29 | Daporinad | 8.371773 |
| HCT-15 | Daporinad | 8.366854 |
| ECC10 | Patupilone | 8.365064 |

| | | |
|---|---|---|
| SNU-C2B | Patupilone | 8.359697 |
| TMK-1 | Patupilone | 8.35942 |
| SW48 | Patupilone | 8.355908 |
| SNU-C5 | Patupilone | 8.354289 |
| TGBC24TKB | Patupilone | 8.348168 |
| HGC-27 | Filanesib | 8.34229 |
| RKO | Ixabepilone | 8.337311 |
| AGS | Elesclomol | 8.32739 |
| HCT-116 | Patupilone | 8.32688 |
| HLE | Ixabepilone | 8.325175 |
| 23132-87 | Daporinad | 8.322321 |
| SNU-175 | Vinflunine | 8.322167 |
| RF-48 | Patupilone | 8.318508 |
| Hep3B2-1-7 | Mipsagargin | 8.317402 |
| SK-CO-1 | Vinblastine | 8.314604 |
| NUGC-4 | Mipsagargin | 8.309074 |
| ECC12 | Bortezomib | 8.306465 |
| TGBC11TKB | Patupilone | 8.304866 |
| MKN1 | Patupilone | 8.303726 |
| HSC-39 | Bortezomib | 8.29594 |
| SNU-1040 | Patupilone | 8.294444 |
| KATOIII | Patupilone | 8.281513 |
| HuTu-80 | Ixabepilone | 8.280867 |
| HGC-27 | Vinblastine | 8.277131 |
| SNU-387 | Patupilone | 8.271601 |
| SNU-81 | Patupilone | 8.269637 |
| NUGC-4 | Patupilone | 8.264868 |
| SK-CO-1 | Vindesine | 8.257339 |
| HLE | Daporinad | 8.251772 |
| SW620 | Ixabepilone | 8.250611 |
| AGS | Bortezomib | 8.247795 |
| HuH-7 | Mipsagargin | 8.247296 |
| SNU-407 | Daporinad | 8.244984 |
| 23132-87 | Bortezomib | 8.241555 |
| HCC2998 | Vinblastine | 8.241009 |
| HuCCT1 | Mipsagargin | 8.239591 |
| JHH-6 | Mipsagargin | 8.239344 |
| 23132-87 | Patupilone | 8.23886 |
| SNU-175 | Deacetoxyvinzolidine | 8.238698 |
| SW48 | Mipsagargin | 8.238303 |
| HGC-27 | Vindesine | 8.236927 |

| | | |
|---|---|---|
| TGBC24TKB | Mipsagargin | 8.235001 |
| SNU-407 | Patupilone | 8.232476 |
| HSC-39 | Filanesib | 8.230596 |
| ETK-1 | Mipsagargin | 8.229173 |
| JHH-1 | Patupilone | 8.225102 |

En yüksek değerli olan 100 çift için değerler gösterilmiştir.

## 8.5. EK-5: Etik Kurul İzin Belgesi



T.C.
HACETTEPE ÜNİVERSİTESİ
Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu

Sayı : 16969557 -1639
Konu :

20.10.2020

Doç. Dr. Tunca DOĞAN
Bilişim Enstitüsü
Sağlık Bilişimi Anabilim Dalı
Öğretim Üyesi

Sayın Doç. Dr. DOĞAN,

Kurulumuza değerlendirilmek üzere sunduğunuz GO 20/964 kayıt numaralı ve "*Yapay Öğrenme Bazlı Hesaplamalı Modelleme ile Geniş Çaplı Kanser Hücre Hattı İlaç Yanıt Tahmini*" başlıklı proje Kurulumuzun 20.10.2020 tarihli toplantısında değerlendirilmiş olup, çalışmada açık erişimli veri tabanından veri seti kullanılacağı anlaşılmıştır. Gönüllü insanlar üzerinde gerçekleştirilecek nitelikte olmayan bu tip çalışmalar Etik Kurulların kapsamı dışında kalmaktadır.

Bu yazı ilgili protokolün etik açıdan incelendiğini belirtilmek için Etik Kurul kararı yerine geçmek üzere hazırlanmıştır.

Prof. Dr. Ayşe Lale DOĞAN
Başkan

EK :
Toplantı Katılım Tutanağı.

ASLI GİBİDİR

Hacettepe Üniversitesi Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu
06100 Sıhhiye-Ankara
Telefon: 0 (312) 305 1082 • Faks: 0 (312) 310 0580 • E-posta: goetik@hacettepe.edu.tr

Ayrıntılı Bilgi için:

## 8.6. EK-6: Tez Çalışması Orijinallik Raporu

turnitin

# Dijital Makbuz

Bu makbuz ödevinizin Turnitin'e ulaştığını bildirmektedir. Gönderiminize dair bilgiler şöyledir:

Gönderinizin ilk sayfası aşağıda gönderilmektedir.

| | |
|---|---|
| Gönderen: | Umut Onur Özcan |
| Ödev başlığı: | Umut Onur Özcan - YL Tez final |
| Gönderi Başlığı: | Lisans üstü tez kontrolü |
| Dosya adı: | a__TEZ_umutoozcan_v9.docx |
| Dosya boyutu: | 19.86M |
| Sayfa sayısı: | 298 |
| Kelime sayısı: | 54,381 |
| Karakter sayısı: | 390,093 |
| Gönderim Tarihi: | 21-Eyl-2022 07:09ÖS (UTC+0300) |
| Gönderim Numarası: | 1905457179 |

T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

YAPAY ÖĞRENME BAZLI HESAPLAMALI MODELLEME İLE GENİŞ ÇAPLI KANSER HÜCRE HATTI İLAÇ YANIT TAHMİNİ

Umut Onur ÖZCAN

Biyoinformatik Programı
YÜKSEK LİSANS TEZİ

ANKARA
2022

Copyright 2022 Turnitin. Tüm hakları saklıdır.

TEZİN TAM BAŞLIĞI : YAPAY ÖĞRENME BAZLI HESAPLAMALI MODELLEME İLE GENİŞ ÇAPLI KANSER HÜCRE HATTI İLAÇ YANIT TAHMİNİ

ÖĞRENCİNİN ADI SOYADI : UMUT ONUR ÖZCAN

DOSYANIN TOPLAM SAYFA SAYISI : 298

## Lisans üstü tez kontrolü

ORİJİNALLİK RAPORU

| %6 | %6 | %4 | %4 |
|---|---|---|---|
| BENZERLİK ENDEKSİ | İNTERNET KAYNAKLARI | YAYINLAR | ÖĞRENCİ ÖDEVLERİ |

BİRİNCİL KAYNAKLAR

| | | |
|---|---|---|
| 1 | www.openaccess.hacettepe.edu.tr:8080<br>İnternet Kaynağı | %1 |
| 2 | academic.oup.com<br>İnternet Kaynağı | <%1 |
| 3 | bmcbioinformatics.biomedcentral.com<br>İnternet Kaynağı | <%1 |
| 4 | www.biorxiv.org<br>İnternet Kaynağı | <%1 |
| 5 | wjgnet.com<br>İnternet Kaynağı | <%1 |
| 6 | medium.com<br>İnternet Kaynağı | <%1 |
| 7 | bmcgenomics.biomedcentral.com<br>İnternet Kaynağı | <%1 |
| 8 | anket.cs.hacettepe.edu.tr<br>İnternet Kaynağı | <%1 |
| 9 | www.frontiersin.org<br>İnternet Kaynağı | <%1 |

# 9. ÖZGEÇMİŞ