

Makine Öğrenmesi Modelleri ile Kansere Hücresel Etkileşim Tahmini

Cancer Cell Line–Drug Interaction Prediction with Machine Learning Models

*Makale Bilgisi / Article Info

Alındı/Received: 26.09.2024

Kabul/Accepted: 04.12.2024

Yayımlandı/Published: 10.06.2025

Berfu MERGEN^{1,*}, Gıyasettin ÖZCAN^{1,2*}

¹ Bursa Uludağ Üniversitesi, Sağlık Bilimleri Enstitüsü, Translasyonel Tıp Anabilim Dalı, Bursa, Türkiye

² Bursa Uludağ Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Bursa, Türkiye



© 2025 The Authors | Creative Commons Attribution-Noncommercial 4.0 (CC BY-NC) International License

Öz

Bu çalışmada, küçük hücreli akciğer karsinomunda farmakogenomik etkileşimler analiz edilmiştir. Bu analiz sonucunda ilaçların mutasyon yüküne bağlı olarak duyarlılığını makine öğrenmesi yöntemleri ile tahmin edilmesini sağlayacak veri toplama, manipülasyon ve model geliştirme süreçleri yapılmıştır. İlk veri kaynağı hücre hatları ve bunların mutasyon bilgilerini içermektedir. İkinci veri kaynağı hücre hatlarına ait detaylı bilgileri içermektedir. Üçüncü veri kaynağı ise ilaç-hücre etkileşimlerini ve hücre hatlarına karşı ilaç duyarlılığını içermektedir. Birleştirilen verilerden farklı mutasyon yük bilgilerinin sayılarak ilaç bileşikleri, hücre hatları, mutasyon yükleri, doku ve IC50 özellikleri tek bir veri kümesinde toplanmıştır. Çalışmanın bundan sonraki aşamasında türetilen ilaç hassasiyeti-mutasyon yükü ilişkisini sunabilecek yeni veri kümesi, mutasyon yüküne göre ilaç direnci etkisi tahmin edilmek üzere makine öğrenmesinde kullanılmıştır. Bu amaçla tahmin için üç farklı makine öğrenmesi algoritması test edilmiştir. Makine öğrenmesi performans analizi için RMSE, R^2 ve MAE sonuçları bulunmuş ve karşılaştırılmıştır. Elde edilen sonuçlara göre geliştirdiğimiz XGBoost makine öğrenmesi modeli hücre-ilaç arasındaki IC50 skorunu anlamlı oranda tahmin etmiştir. Bu sayede ilaçların mutasyonlara direncine ve etkisine dair ön bilgi sunulmaktadır. Bunun yanı sıra çalışmada hangi mutasyon türlerinin nicel sayısının ilaç direncinde daha fazla etki gösterdiğini makine öğrenmesi analizleri ile sunulmuştur.

Abstract

In this study, we addressed pharmacogenomic interactions in lung small cell carcinoma. For this purpose, data collection, data manipulation and machine learning algorithms were utilized. By combining three open-source datasets, a new dataset is generated. The first data source contains cell lines and their mutation information. The second data source contains detailed information about the cell lines. The third dataset contains drug-cell interactions and drug sensitivity of cell lines. By combining the utilized data sources, a new dataset was obtained by counting different mutation load information. Thus, chemical compounds, cell lines, mutation loads, tissue and IC50 characteristics were collected in a single dataset. In the next stage of the study, the new dataset that could present the derived drug sensitivity-mutation load relationship was used in machine learning to predict the drug resistance effect according to mutation load. For this purpose, three different machine learning algorithms were tested for prediction. For machine learning performance analysis, RMSE, R^2 and MAE results were found and compared. According to the results, the XGBoost machine learning model we developed significantly predicts the IC50 score between cell-drug. In this way, it provides preliminary information on the extent to drug resistance and drug effect. In addition, the study presents which mutation types have a greater effect on the quantitative number of drug resistance through machine learning analysis.

Anahtar Kelimeler İlaç Duyarlılığı Tahmini; Mutasyon Yükü; Makine Öğrenme; Kişiselleştirilmiş Tıp; XGBoost

Keywords Drug Sensitivity Prediction; Mutational Burden; Machine Learning; Personalized Medicine; XGBoost

1. Giriş

Son yıllarda, hassas tıbbın artan ilgi ve büyük ölçekli farmakogenomik veri kümelerinin mevcudiyeti nedeniyle ilaç duyarlılığının tahmin edilmesine ilgiyi artırmıştır ve ilaç duyarlılığının test edilmesi rutin bir prosedüre dönüşmeye başlamıştır (Vis vd. 2016). Farmakogenomik, hassas tıpta ortaya çıkan, ilaç seçimini ve dozajı hastanın genetik özelliklerine uygun hale getiren yaklaşımlardan biridir. Son yıllarda, uluslararası bilimsel konsorsiyumlar tarafından çeşitli farmakogenetik kılavuzlar yayınlanmıştır, ancak klinik uygulamada kabulü hala

zayıftır (Cecchin ve Stocco 2020). Farmakogenomik alanı genetik mekanizmaları tanımlayarak ilacın etkinliğini artırmak ve toksisiteyi azaltmak için genetik testlerin uygulanmasını araştırır. Farmakogenomik, değişken ilaç yanıtının genetik temelini anlamak için bir araç olarak kullanılabilir. Mevcut ilaçların kullanımını yeni endikasyonlara genişletmek ve yeni ilaçlar geliştirmek için de nispeten yeni bir araştırma alanıdır (Roden vd. 2019). Bu amaçla gen mutasyonlarına ait veriler kullanılarak belirli bir ilacın etkinliğinin anlaşılması hedeflenmektedir. Böylece beşinci yüzyıldan beri var olduğu bilinen

kişiselleştirilmiş tıp konusunda yeni bir aşamaya erişilmiştir (Steele 2009). Kişiselleştirilmiş tıp hasta için öngörülen tedavi yanıtına veya hastalık riskine göre tıbbi kararların alındığı, bu strateji ile hastalığı değil, hasta bireyleri farklı gruplara ayıran bir tıbbi modeldir. En kısa tanımı ile tıbbi tedavinin her hastanın bireysel özelliklerine göre uyarlanmasıdır. İlaç duyarlılığı modellemesinin doğruluğu ve hassasiyeti büyük ölçüde veri türleri, özellik seçimi, model seçimi ve model doğrulama gibi ilişkili faktörlere bağlıdır (Matlock vd. 2018).

Makine öğrenmesi, çeşitli öğrenme teknikleri ve bu tekniklerin oluşturduğu algoritmaları kullanarak bilgisayarların mantıklı sonuçlar üretmesini sağlar. Diğer yapay zeka uygulamalarından farklı olarak, makine öğrenmesi, bilgisayara belirli kurallar vererek çıktıyı değiştirmek yerine, kuralların istatistik kullanarak kendisi tarafından oluşturulması gerektiği anlamına gelir (Hastie vd. 2009). Bu seviyeye ulaşabilmesi için bilgisayara çeşitli veri girdileri verilir ve daha sonra en uygun öğrenme algoritmalarıyla çıktı oluşturulur. Yapay zeka, bilgisayarların insan aklını taklit ederek sorunları çözmesine olanak veren uygulamaların genel ismidir.

Klinikte kanser tedavisinde kullanılacak ilaç seçiminde, hastanın kanser hücresi özelliklerinin belirlenmesiyle elde edilen bilgi; tümör alt popülasyonları için en uygun tedavinin seçilmesine ve potansiyel olarak etkisiz tedavilerden kaçınılmasına olanak tanıyabilmektedir. Genetik özellikler ve ilaç yanıtı arasındaki ilişkiler, ilaç etki mekanizmaları hakkında değerli yeni bilgiler sağlayabilir. Literatürdeki kanser hücre hattı verilerinin yeni bilgisayar teknolojileri yardımıyla taranması gelecekteki çalışmalara yön verebilecektir. Onkoloji ilaç kombinasyonları terapötik yanıtları iyileştirebilir ve hastalar için tedavi seçeneklerini artırabilir. Olası kombinasyonların sayısı çok fazladır. Sistemik taramalar, tanımlanmış hasta alt tiplerinde klinik açıdan anlamlı, eyleme dönüştürülebilir ve klinikte kullanılabilir kombinasyonları belirleyebilir (Bashi 2024).

Kanser biyolojisini incelemek ve çeşitli kanser tedavilerini test etmek için en popüler yöntem hücre hatlarının kullanılmasıdır. Kanser hücre hatlarının türetildikleri orijinal kanser türünün özelliklerini (gen ifadesi, ilaç duyarlılığı, genotipik ve fenotipik özellikler ve mutasyon durumu) sergilemesi beklenmektedir (Langdon 2004, Masters 2000). Bununla birlikte, laboratuvar ortamı ve hücre hattında diğer normal hücre tiplerinin bulunmaması, bir hücre hattının in vivo kanser büyümesini taklit etme yeteneğini ve anti-kanser ilaçlarının etkinliğini kısıtlar (Gillet vd. 2011).

İnsan DNA'sındaki gen varyantları son yirmi yılda yoğun bir şekilde incelenmiş ve hastalıklara karşı tedavi stratejileri araştırılmıştır. Bu amaçla tedaviye verilen klinik yanıtlar ve ilaçlara verilen yanıtları sunan biyobelirteçler araştırılmaktadır. Bu bağlamda açık kaynaklı Genomics of Drug Sensitivity in Cancer Database (GDSC) araştırmacıların kullanımına sunulmuştur (Yang vd. 2013).

GDSC, kanser hücreleri hakkında ilaç duyarlılığı ve ilaç yanıtları konusunda açık kaynaklı verilerden oluşmaktadır. Bu deneysel veriler, araştırmacıların ilaç-kanser etkileşimlerini silico veriler aracılığıyla incelemesine olanak sunmaktadır. GDSC veritabanı incelendiğinde ilaçlar, hedef gen, hücre hatları, dokular, kanser özellikleri ve yarı maksimal inhibitör konsantrasyon (IC50) arasındaki ilişki araştırılabilir. Ayrıca ilaçlar ve bileşikler gen ile etkileşime girdiğinde gen ürünlerini artırabilirler veya inhibe edebilirler. Gen ürünü bir genin ifadesi sonucu elde edilen biyokimyasal (DNA ya da protein) maddedir. Elde edilen gen ürünü miktarı zaman zaman ilgili genin niteliğini anlamak için de kullanılmaktadır. Aşırı miktardaki gen ürünü kansere yol açan onkogen etkinliği gibi hastalık yapıcı alellerle ilişkilendirilmektedir (Fearon ve Vogelstein, 1990).

Literatürde ilaç hassasiyet ölçüsü olarak hücre yanıt eğrisinin tamamını kullanmak yerine, hassasiyeti temsil etmek için eğrinin IC50 (hücre canlılığını ilk yanıtın %50'sine düşürmek için gereken ilaç konsantrasyonu) gibi belirli bir özelliği kullanılır. IC50'ler genellikle logaritmik bir eşleme fonksiyonu kullanılarak sıfır ile bir arasındaki hassasiyet değerlerine dönüştürülür (Berlow vd. 2014). Yakın literatürde farmakogenomik konusunda yapay zeka çalışmaları hızla artmaktadır. (Wang vd. 2020) farmakogenomik verilerden ilaç ve hücre hattı özelliklerini ortaklaşa öğrenmek için sadece grafik modeli değil, aynı zamanda tahmin sonuçlarının korelasyonunu güçlendirmek için ilaç ve hücre hattı benzerlik ağı modelini de içeren çift katmanlı güçlendirilmiş işbirlikçi konu regresyonu (DS-CTR) kullanılarak doğru ilaç duyarlılığı tahmini üzerine bir yenilik modeli tanıtılmıştır.

İlaç direnci veri tabanlarını analiz ederek ilaç direnci mutasyonlara bağlı olarak öngörüsü yapan makine öğrenmesi modelleri geliştirmek yakın zamanda literatürde yaygınlaşmaktadır (Sharma ve Rani 2020, Gao vd. 2021, Singh ve Kaushik 2023). Zira bazı mutasyonlar iyi huylu katkılar sağlarken, diğer bazı mutasyonlar patojenik olabilir ve kanser gelişimine katkıda bulunabilir.

Farmakogenomik veri tabanları son zamanlarda tahmin modelleri için eğitim verisi olarak kullanılmaktadır. (Barretina vd. 2012, Yang vd. 2013). İlaç yanıtı tahmin

modellerini kategorize etmek için iki yaklaşım kategorisi kullanılabilir:

- (i) tek veya çoklu-omik tabanlı yaklaşımlar (Bomane vd. 2019, Chiu vd. 2019, Costello vd. 2014, Wei vd. 2019)
- (ii) ilacın kimyasal özellikleriyle desteklenen yaklaşımlar (Park vd. 2022)

Daha önceki araştırmalara göre, bir derin öğrenme modelinin doğruluğu, içerdiği katman sayısı veya derinlik ile artmaktadır (Lecun vd. 1998, Vestias 2020).

Bir çalışmada, değişen veri kümesi boyutları ve omics tipi kullanımlar gibi çeşitli veri kümesi koşulları altında IC50 değerlerini tahmin etmede regresyon tabanlı ML ve DL modellerini kapsamlı bir şekilde değerlendirmişlerdir. Çoğu durumda, DL modelleri çeşitli koşullarda ML modellerinden daha iyi performans göstermiştir (Park vd. 2022). Bir çalışmada, farklı ilaç kimyasal parmak izleri ve mutasyon durumları gibi 27.756 özellik kullanılarak üç ilaç yanıt sınıfını tahmin etmek için DeepIC50 adı verilen 1 boyutlu bir konvolüsyonel sinir ağı modeli geliştirilmiştir. Sonuç olarak DeepIC50, pan-kanser hücre hatlarında hücre canlılığı ve IC50 tahmin doğruluğu açısından iki ayrı kanser hücre hattı veri kümesinden daha iyi performans göstermiştir. İlaç keşif sürecinde DeepIC50, yeni bileşiklerin çeşitli kanser hücre hatlarının duyarlılığını nasıl etkileyeceğini doğru bir şekilde tahmin edebilir (Joo vd. 2019).

Kanser Genom Atlası (TCGA) gibi çok büyük klinik kanser genomik veri setlerinde, atanan ilaç yanıtı çok önemlidir. Bu yöntem, kanser hücre hatlarının büyük panellerinde tedavi yanıtını gen ifadesine bağlayan istatistiksel modeller oluşturmayı ve ardından bu modelleri klinik veri setlerinden (TCGA gibi) tümör gen ifadesi verilerini analiz etmek için kullanmayı içerir. Bu, her hastanın ilacı için bir farmakolojik yanıt üretir. Daha sonra, bu ilaç yanıt verileri ölçülen somatik genetik varyasyonlarla ilişkilendirilir (Geeleher vd. 2017).

İlaç direnci öngörüsü yapmak için kullanılan algoritmalar makine öğrenmesi temellidir. Makine öğrenmesinde amaç veriden öğrenme yapmaktır (Hastie vd. 2009). Literatürde birçok makine öğrenmesi algoritması mevcuttur. SVM (Support Vector Machine) algoritması verileri sınıflandırabilmek için uzayı çok boyutlu ayrıç denklemleri ile ifade etmektedir (Boser vd. 1992). SVM doğru tahminleme ve hız açısından verimli bir sonuçlar sunmaktadır. Öte yandan aynı veri üzerinde birden fazla sınıflandırıcı tanımlamak bir başka alternatif olmuştur. Bu yöntemde üretilen sınıflandırıcıların hata performanslarından ders alınarak bir sonraki

sınıflandırıcının hesaplanması amaçlanmaktadır. XGBoost ve CatBoost bu sınıflandırma yönteminin bilinen örnekleridir (Chen ve Guestrin 2016, Prokhorenkova vd. 2018).

Yarı Maksimal İnhibitör Konsantrasyon veya IC50, bir maddenin (ilaç, inhibitör veya diğer biyoaktif bileşenler gibi) belirli bir biyolojik veya biyokimyasal işlevi inhibe etme gücünü gösteren önemli bir ölçümdür. Bu, farmakoloji ve biyokimya alanında çok önemlidir. IC50, bir maddenin bir biyolojik süreci (örneğin enzim aktivitesi, reseptör bağlanması veya hücre büyümesi gibi) %50 oranında inhibe etmek için gereken konsantrasyonunu gösterir. Aynı inhibisyon seviyesine ulaşmak için daha az maddeye ihtiyaç duyulduğundan, daha düşük IC50 değeri bileşiğin daha güçlü olduğunu göstermektedir. IC50 potansiyel terapötik ajanların etkinliğini ve güvenlik profilini değerlendirmek, bunları diğerleriyle karşılaştırmak ve dozaj stratejilerini geliştirmek için önemli bir araçtır. Farmakolojide standart bir ölçüttür ve umut vadeden bileşiklerin laboratuvarından klinik testlere taşınmasına yardımcı olur (Berrouet vd. 2020).

Bu çalışmada akciğer küçük hücreli karsinomunda mutasyon yükü analizi yapılmıştır. Bu amaçla Sanger Enstitüsüne ait üç farklı veri kümesi toplanıp birleştirilerek mutasyon yük analizi için tek bir veri kümesi türetilmiştir. Türetilen veri kümesinde küçük hücreli akciğer karsinomu ile ilişkili olan veriler filtrelenerek analize uygun bir veri kümesi elde edilmiştir. Elde edilen veriler ile ilaçların hangi mutasyonlara karşı daha duyarlı olabileceğine dair analizler yapılmıştır. İlaç duyarlılığı ile mutasyonlar ve mutasyon yükleri arasındaki ilişkiyi belirlemek için makine öğrenmesi modelleri geliştirilmiştir. Filtrelenerek elde edilen tüm veriler arasında Cell Model Password sitesinde yer alan All Mutations verisi indirilmiştir (İnt. Kyn. 1). Bunun yanı sıra hücre hatlarına dair veri ve hücre hatlarının bilgileri farklı kaynaklardan alınarak birleştirilmiştir (İnt. Kyn. 2, İnt. Kyn. 3). Geliştirilen SVM Boser vd. (1992), CatBoost Prokhorenkova vd. (2018) ve XGBoost Chen ve Guestrin (2016) modelleri ile ilaç etkileri tahmin edilmesi amaçlanmıştır. Geliştirilen modeller arasında CatBoost modelinin mutasyonların ve toplam yükünün belirli bir ilaca hassasiyetini tahmin edebilme ve potansiyel ilaç adaylarını sunma konusunda destek sunma potansiyeli olduğunu göstermektedir.

2. Materyal ve Metot

Çalışmanın ilk safhasında farklı veri kümelerinden veriler toplanarak ilaç hassasiyeti-mutasyon yükü ilişkisini sunabilecek veri hazırlanmıştır. Zira mutasyon yükünün gereksinim duyulan formatta doğrudan sunan hazır veri kaynağı bulunamamıştır. Bunun yerine farklı

kaynaklarından toplanan veriler kullanılarak mutasyon yükünü tanımlamak mümkün olmuştur.

2.1 Veri Toplama

Hücre ve mutasyon bilgilerinin ilaç direncine etkisi konusunda sunulan açık kaynaklardan GDSC' den Yang vd. (2013) veri seti kaynakları toplanmıştır. Elde edilen kaynaklar arasında hücre hatlarında yer alan mutasyonları tanımlayan tüm Mutasyonlar verisi Cell Model Passports üzerinden indirilmiştir (van der Meer *et al.* 2019). Bu kaynakta gene_id, Hücre_Model_id, mutasyonun driver-kodlayıcı olup olmadığı, etkisi, vaf skoru yer almaktadır. Öte yandan bu kaynakta ilgili hücre hattının hangi kanser türlerine maruz kaldığı yer almamaktadır.

Bunun yanı sıra, ilaçların belirli mutasyonlu hücrelere etkisini analiz etmek önemlidir. İlaçların ilgili mutasyonlara sahip hücrelere etki oranlarını hesaba katmak için ek kaynaklara gereksinim duyulmuştur.

GDSC veri tabanında yer alan IC50 verisi, çok sayıda ilacın hücre hatlarına verdiği ilaç tepkisini sunmaktadır. İlaçlar ve ilgili veriler GDSC açık kaynaklı veri tabanından elde edilmiştir. Bu veri sayesinde hücre hattı ile ilaç etkisi arasındaki ilişki daha iyi analiz edilebilecektir. Verinin toplandığı gün itibarı ile hücre hatlarının 265 ilaca veya aday ilaca karşı etkisi sunulmuştur.

GDSC veri tabanında yer alan hücre listesi verisi her bir hücre hattının özelliklerini ve maruz kaldığı kanser türlerini sunmaktadır. Bunun yanı sıra mutasyon verilerini sunan veri ile IC50 verisi arasında ortak anahtar barındırması nedeniyle önemi daha da çok artmaktadır.

Toplanan üç veri kaynağı Hücre Listesi verisi anahtar kullanılmak üzere birleştirilmiştir. Öte yandan birleşme sonucu ortaya çıkan verinin çok sayıda satır içermesi nedeniyle filtreleme yapılması elzem olmuştur. Çalışmada lung_small_cell_carcinoma verilerinin tercih edilmesinin nedeni açık kaynaklı veri olarak diğer kanser türlerine göre küçük hücreli akciğer kanseri verilerinin sayıca fazla olmasıdır.

2.2 Lung Small Cell Carcinoma verisi

GDSC, farklı kanser türlerine karşı ilaç hassasiyeti konusunda veri sunmaktadır. Bu çalışmada sadece lung_small_cell_carcinoma verilerine ve ilgili ilaç hassasiyetine odaklanılmıştır.

Filtreleme yapmak üzere subtype özelliği "lung_small_cell_carcinoma" olan veriler seçilmiştir. Filtreleme sayesinde veri boyutu küçülmüş ve hesaplanabilir boyuta indirgenebilmiştir. Filtreleme işlem sonucu elde edilen mutasyon verisi 512464 satırdan oluşmaktadır.

2.3 Makine Öğrenmesi için Veri Seti Oluşturma

Bir sonraki adımda makine öğrenmesinde kullanıma uygun verinin elde edilmesi sağlanmıştır. Bu amaçla Lung Small Cell Carcinoma verisinde yer alan mutasyonlara dair bazı özellikler sayılarak toplanmıştır. GDSC veri kaynaklarından üretilen veri setinde her ilaç için ayrı ayrı olmak üzere aşağıdaki listede belirtilmiştir:

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

2.4 Makine Öğrenmesi Uygulaması

Sanger Enstitüsü kaynakları kullanılarak beş farklı ilaç için yukarıda belirtilen tüm özellikler toplanmıştır. İlaç seçimi yapılırken ilaç-hücre etkileşiminde IC50 varyasyonunun geniş tutulabilmesi amaçlanmıştır. Bunun yanı sıra ilgili hücrenin küçük hücreli akciğer karsinomu olmasına dikkat edilmiştir. Üretilen veride her bir ilaç için; ilaç-hücre-ilaç hassasiyeti skorları yer almaktadır. Her bir hücre için yukarıdaki listede sıralanan özelliklerin yer alması sağlanmıştır. Bu sayede her hücrede yer alan ve sayılan mutasyon yüklerine bağlı olarak ilaçların hücrelere olan etkileri değerlendirilmek amaçlanmıştır. Her bir satırda bu beş ilacın belli bir hücre için ilaç direnci yer almaktadır. Çalışmada lung_small_cell_carcinoma tanısı konulmuş hücreler için ilaç analizleri yapılmıştır. Beş ilacın verileri birleştirilerek bir dosyada toplanmıştır. 241 ilaç-hücre-ilaç hassasiyeti ilişkisini analiz edebilmek amacıyla 14 sütunlu veri oluşturulmuştur.

2.5 Makine Öğrenimi Uygulaması

Üretilen 241 satır ve 14 özellikten oluşan veri ile makine öğrenmesi uygulaması yapılmıştır. Bu özellikler arasında İlaç Direnci bağımlı değişken olarak tanımlanmış ve tahmin edilmiştir. Tahminleme sonucunda hedeflenen ilaç ve kanserli hücre bilgisine bağlı olarak ilaç direncinin tahminini yapmaktır. Çalışmayı özgün kılan nokta ise farklı mutasyon yüklerinin toplamının ilaç direncine etkisinin

incelenmesidir. Elde edilecek sonuçlar doğrultusunda mutasyon yüküne ve mutasyon yük özelliğine bağlı olarak ilaç direncinin önceden tahmin edilmesidir. SCLC için IC50 değeri 61 farklı hücrenin yaklaşık 500.000 olası mutasyonu dikkate alınarak analiz edilmiştir. Mutasyon verilerinin bir kısmı eğitim için kullanılmıştır. Eğitimden sonra IC50 değeri diğer bazı hücrelerin mutasyonlarına göre tahmin edilmeye çalışılmıştır. Analiz edilen ilaçlar için %60 R2 tahmin değerine ulaşılmıştır. Makine öğrenmesi sürecinde XGBoost, CatBoost ve SVM algoritmaları denenmiş veriye uyumlu en iyi modelleri elde etmek için yazılım tabanlı deneysel çalışmalar yapılmıştır. XGBoost, Catboost ve SVM modellerinin geliştirilmesi sırasında üretilen verilerin %80'i eğitim, kalanı ise test amacıyla kullanılmıştır. Ayırma işlemi yapılırken hücre ismi özelliği dikkate alınmıştır. Böylece her hücre için yeteri kadar test verisinin ayrılmış olması emniyete alınmıştır. Bu doğrultuda elde edilen verilerin adil bir şekilde %80'i eğitim için ayrılmıştır.

2.6 İdeal Model Parametre Belirlenmesi

CatBoost algoritması için kullanılan parametreler Çizelge 1'de sunulmuştur. Bir diğer makine öğrenmesi modeli olan XG Boost algoritması ise benzer şekilde Çizelge 2'de belirtilen başlangıç parametreleri ile çalıştırılmıştır.

Çizelge 1. CatBoost algoritması için kullanılan parametreler

Parametre	Değer
Seed	13
İterasyon	300
Learning rate	0.1
Derinlik	6
Loss function	RMSE
eval metric	RMSE

Çizelge 2. XG Boost algoritması için kullanılan parametreler

Parametre	Değer
max depth	6
nrounds	300
verbose	0
Learning rate	0.1

Son olarak SVM algoritmasına ait model denemeleri uygulanmıştır. SVM modelinde potansiyel en iyi modeli bulabilmek için C ve σ parametreleri denenmiştir. Bunun yanı sıra 10-fold cross validation kullanılmıştır. SVM deneylerinde test edilen C ve σ parametreleri aşağıdaki gibidir:

$$C = c(0.25, .5, 1),$$

$$\sigma = c(0.1, 0.01, 0.001)$$

Elde edilen sonuçlar, her algoritma için farklı parametre deneylerinin performans açısından en iyisi araştırılarak

elde edilmiştir. Tüm uygulama kodları R programı ile yazılmıştır. Program icrası için 7. nesil i7 işlemcili ve 16 GB belleğe sahip bilgisayar kullanılmıştır.

3. Bulgular

Deney kısmında ilk önce mutasyon yük verisi üretme ve sonuçları anlatılmaktadır. İkinci kısımda ise üretilen veriler ile ilaç direncinin mutasyon yükü bilgilerine bağlı olarak tahmin edilmesi amaçlanmıştır. Bu amaçla makine öğrenmesi modelleri geliştirilmiş ve performansı RMSE, Rsquared ve MAE cinsinde ölçülerek kıyaslanmıştır.

3.1 Mutasyon Yük Veri Üretmenin Zorluğu

Hücrelerde yer alan tüm mutasyonları barındıran dosya 893 MB büyüklüğündedir. Hücrelere dair kapsamlı bilgileri içeren Hücre Listesi dosyası ise 128 KB boyutundadır. Bu iki dosyanın kartezyen çarpımı ile birleştirilmesinin bellek yetersizliğine neden olacağı aşikardır. Bu nedenle bu işlemde sonra sadece küçük hücreli akciğer kansinomu verileri filtrelenerek veri küçültülmüştür. İlaçlar ile hücreler arasındaki hassasiyet skorlarının saklandığı IC50 dosyası ise 3.160 KB boyutundadır. Daha sonra filtrelenmiş dosya ile ilaçların IC50 bilgilerinin olduğu dosyanın birleştirilmesi hedeflenmiştir. Birleştirilen dosyada hem seçilen ilaçlar için IC50 değerleri hem de hücre bilgileri yer almaktadır. Öte yandan ham verilerin 8GB RAM'a sahip bilgisayarda birleştirilip filtrelenmesinin RAM yetersizliği nedeniyle gerçekleşmediği görülmüştür. Yüksek RAM gereksinimi nedeniyle veri manipülasyonu esnasında filtreleme işlemin mümkün mertebe öncelikte tercih edilmesi gerekmektedir.

3.2 Makine Öğrenme Algoritma ile Tahminler

Kanserli hücre ve ilaç arasındaki IC50 değerini tahmin etmek için Makine Öğrenme algoritma performansları incelenmiş ve sonuçlar aşağıda değerlendirilmiştir.

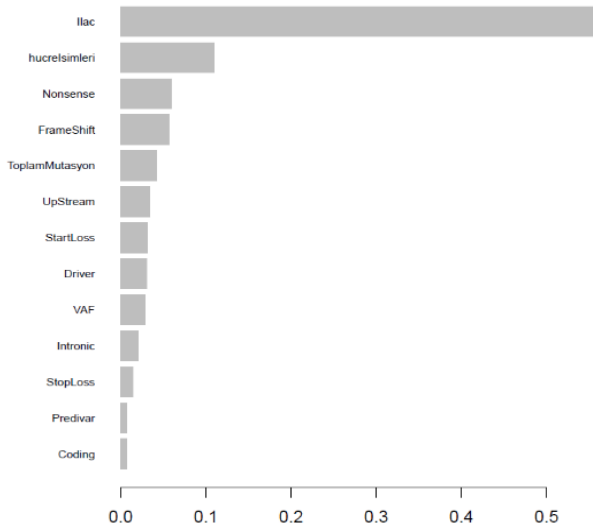
XGBoost, CatBoost ve SVM algoritmaları baz alınarak geliştirilen modeller aşağıdaki Çizelge 3'te sunulmuştur.

Çizelge 3. XGBoost, CatBoost ve SVM algoritmaları baz alınarak geliştirilen modeller

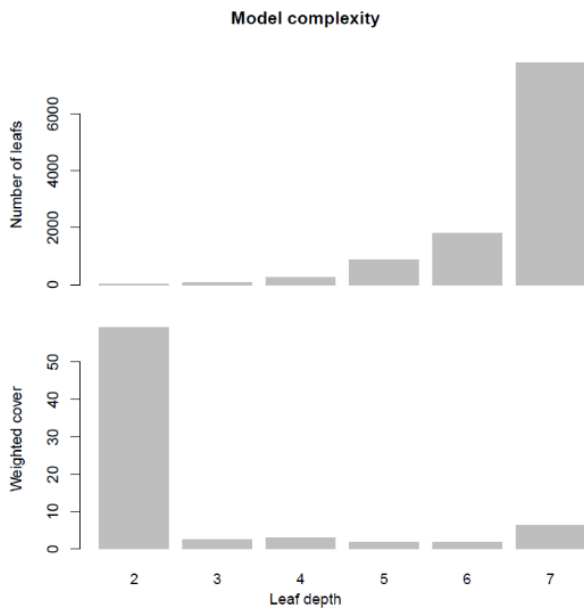
Model	RMSE	Rsquared	MAE
XGBoost	1.602683	0.6045291	1.231046
CatBoost	1.654969	0.5796071	1.314209
SVM	2.47293813	0.04089099	1.96588405

Tabloda görüldüğü üzere en başarılı RMSE, R2 ve MAE skorlarını XGBoost algoritması sunmuştur. Tabloda belirtilen sonuçlar doğrultusunda Gradient temelli boosting algoritmalarının ilaç hassasiyetini tahmin etme konusunda başarılı sonuçlar sunduğu görülmektedir. Potansiyel tüm ilaçların kanserli hücre etkileşimini deney yapmadan 1.60 RMSE, 0.6 R² ve 1.23 MAE oranında doğru

tahmin edilmesi, mutasyon durumuna göre hangi ilaçların analiz yapılması gerektiği konusunda ön bilgi sunmaktadır. Bu başarımlar oranı bu nedenle önemlidir. XGBoost modelinin tahmin başarımı yaparken hangi özellikleri dikkate alındığına dair bilgi Şekil 1’de sunulmuştur. Verilerin %80’inin eğitimde kullanılması nedeniyle tahmin esnasında ilaç ve hücre isimlerinin yer alması beklenmektedir. Bunların ötesinde en etkili faktörlerin sırasıyla Nonsense, Frameshift, ToplamMutasyon, Upstream ve Start Loss mutasyon sayılarının Driver mutasyon sayısından daha önemli olduğuna ortaya koymaktadır. Bu nedenle bu mutasyonların niceliksel olarak fazla olması durumunda ilaç hassasiyetinin gelişme olasılığının artması ihtimali göz ardı edilmemelidir.



Şekil 1. XGBoost Model grafiği



Şekil 2. Yaprak derinliğine bağlı olarak XGBoost ağacında ağırlıklı kapsam ve yaprak sayısı grafikleri

3.3 Mutasyon Türlerinin Etkisi ve Model Karmaşıklık Analizi

XGBoost algoritmasının yapısına dair teknik özellikler:

Elde edilen sonuçlar dikkate alındığında XGBoost Makine öğrenmesi modelinin yapısı hakkında daha fazla bilgi sunmak ta mümkündür. Tahmin performansını artırmak üzere XGBoost Model karmaşıklık grafiği Şekil 2’de sunulmuştur.

Şekilde yaprak derinliğine bağlı olarak XGBoost ağacında ağırlıklı kapsam ve yaprak sayısı grafiği sunulmuştur. Weighted cover ile ağdaki önemli toplulukları belirlemek amaçlanmıştır. Şekilde 2. ve 7. ağaç derinliklerinin sınıflandırmada etkili olduğu görülmektedir. Bu nedenle derinliğin 7' ye çıkmasının sınıflandırmaya katkı sağladığı izlenmektedir.

Bu çalışmada, ilaç duyarlılığını doğru bir şekilde tahmin etmek için bir yöntem geliştirildi. Bir ilacın kanser hücresi canlılığının yarı maksimal inhibitör konsantrasyonu (IC50), gücünün bir ölçüsü olarak yaygın şekilde kullanılmaktadır. Dolayısıyla daha yüksek ilaç gücü, hücre canlılığı açısından daha düşük IC50 değerleriyle sonuçlanır.

Kanser hücresi canlılığının IC50 değerleri, hücrelerin tümör mutasyon yükü ile tahmin edilmiştir. Tümör mutasyon yükü, kanser hücrelerinin DNA'sında bulunan değişikliklerin (genetik mutasyonlar olarak adlandırılır) sayısıdır. Bu ilaç yanıtı tahmin muammasını ele almak için makine öğrenimi (ML) ve ağ biyolojisi yaklaşımları kullanılmıştır. Bir ilacın yanıtını tahmin etmek, tıbbi etkili bir şekilde uygulamak için çok önemlidir. Araştırmacılar farklı kanser hastalarında ilaç yanıtını değerlendirmek için sıklıkla kanser klinik çalışmalarından elde edilen tedavi sonuçlarına bakmaktadır; ancak bu etik ve mali konular nedeniyle sorunludur. Bunun yerine, kanser hücre dizilerinin genomlarını girdi olarak alan çok sayıda araştırma, kanser hücre dizilerindeki farmakogenomik bilgileri kullanarak farklı ilaçların ilaç yanıtı tahmin modellerini oluşturmaktadır. Bu modeller ile hastaların farklı ilaçlara nasıl yanıt vereceğini tahmin etmek için hastalar ve ilaçla ilişkili hücre hattı arasındaki genomik profilleri karşılaştırmak için kullanılması mümkündür.

4. Sonuçlar ve Tartışma

Bu çalışmada kanser hücresi mutasyon yükünün ilaç direncine etkisi araştırılmıştır. Bu amaçla açık kaynak verileri derlenip farklı mutasyon özellikleri sayılarak yeni bir mutasyon veri yükü seti oluşturulmuştur. Mutasyon sayımı yapılarak üretilen veride ilaç-küçük hücreli akciğer karsinom hücresi-ilaç direnci verilerinin yanı sıra hücredeki farklı mutasyon yüklerinin sayısı da yer almaktadır.

Mutasyon yük bilgilerine göre ilaç hassasiyetinin tahminini yapmak için farklı makine öğrenmesi yöntemleri denenmiş ve farklı modeller geliştirilmiştir. Carboost, XGBoost, SVM ve TabNet algoritmaları kullanılarak geliştirilen modeller ilaç hassasiyetini en çok %60 üzerinde doğru tahmin edebilmiştir. Çalışmanın bir diğer amacı mutasyon yükünün belirlenmesinde mutasyon türünün etkisinin araştırılmasıdır. Kullanılan veri seti için XGBoost makine öğrenmesi tabanlı makine öğrenme modelimiz Nonsense, Frameshift, ToplamMutasyon, Upstream ve Start Loss, ve Driver mutasyonların ilaç hassasiyeti konusunda daha fazla belirleyici olduğu sonucuna varmıştır.

Unutulmamalıdır ki hücre hattındaki bir mutasyon türünün niceliksel olarak fazla olması da ilgili mutasyonun ağırlığının artmasına etki edebilmektedir. Bu nedenle veri kümesinin ve dolayısıyla farklı hastalık kapsamının aynı anda değerlendirilmesi daha etkili sonuçlar sunabilecektir.

Doğru tahmin edilen ilaçların daha kapsamlı deneylere tabi tutulma olanağının sunması nedeniyle sonuçlar anlamlıdır ve potansiyel ilaçların bulunmasına katkı sağlama şansı sunmaktadır. Bundan sonraki çalışmalarda mutasyon yükünün ve özelliklerinin ileriki çalışmalarda daha kapsamlı olarak ele alınmasının önemi ortaya çıkmaktadır.

Etik Standartlar Bildirgesi

Bu çalışma Doç.Dr. Gıyasettin ÖZCAN danışmanlığında Berfu MERGEN tarafından devam etmekte olan Doktora tez çalışmasının bir ürünüdür. Çalışmada kullanılan verilerin tamamı açık kaynaklıdır ve ilgili bilimsel çalışmalara bu makalede atıf yapılmıştır.

Yazarlar tüm etik standartlara uyduklarını beyan ederler.

Yazarlık Katkı Beyanı

Yazar 1: Kaynaklar, Araştırma, Deney, Yazma – original taslak

Yazar 2: Araştırma, Deney tasarımı, Proje Yöneticisi, Yazma –orijinal taslak

Çıkar Çatışması Beyanı

Yazarların bu makalenin içeriğiyle ilgili olarak beyan edecekleri hiçbir çıkar çatışması yoktur.

5. Kaynaklar

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M.F., Monahan, J.E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., Mapa, F.A., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I.H., Cheng, J., Yu, G.K., Yu, J., Aspesi, P. Jr., de Silva, M., Jagtap, K., Jones, M.D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R.C., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J.P., Gabriel, S.B., Getz, G., Ardlie, K., Chan, V., Myer, V.E., Weber,

B.L., Porter, J., Warmuth, M., Finan, P., Harris, J.L., Meyerson, M., Golub, T.R., Morrissey, M.P., Sellers, W.R., Schlegel, R., Garraway, L.A. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**(7391), 603-7.

<https://doi.org/10.1038/nature11003>

Bashi, A.C., Coker, E.A., Bulusu, K.C., Jaaks, P., Crafter, C., Lightfoot, H., Milo, M., McCarten, K., Jenkins, D.F., van der Meer, D., Lynch, J.T., Barthorpe, S., Andersen, C.L., Barry, S.T., Beck, A., Cidado, J., Gordon, J.A., Hall, C., Hall, J., Mali, I., Mironenko, T., Mongeon, K., Morris, J., Richardson, L., Smith, P.D., Tavana, O., Tolley, C., Thomas, F., Willis, B.S., Yang, W., O'Connor, M.J., McDermott, U., Critchlow, S.E., Drew, L., Fawell, S.E., Mettetal J.T., Garnett, M.J., 2024. Large-scale Pan-cancer Cell Line Screening Identifies Actionable and Effective Drug Combinations. *Cancer Discov*, **14**(5), 846-865. <https://doi.org/10.1158/2159-8290.CD-23-0388>.

Berlow, N., Haider, S., Wan, Q., Geltzeiler, M., Davis, L.E., Keller, C., Pal, R. 2014. An Integrated Approach to Anti-Cancer Drug Sensitivity Prediction. *IEEE/ACM Trans Comput Biol Bioinform*. **11**(6), 995-1008. <https://doi.org/10.1109/TCBB.2014.2321138>.

Berrouet, C., Dorilas, N., Rejniak, K.A., Tuncer, N. 2020. Comparison of Drug Inhibitory Effects ([Formula: see text]) in Monolayer and Spheroid Cultures. *Bull Math Biol*. **82**(6), 68. <https://doi.org/10.1007/s11538-020-00746-7>

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

Boser, B.E., Guyon, I.M., Vapnik, V.N. 1992. A training algorithm for optimal margin classifier. Proceedings of the 5th ACM Workshop Pennsylvania, USA. 144–152, <https://doi.org/10.1145/130385.130401>

Cecchin, E. and Stocco, G., 2020. Pharmacogenomics and Personalized Medicine. *Genes (Basel)*, **11**(6), 679. <https://doi.org/10.3390/genes11060679>

Chen, T. and Guestrin, C., XGBoost. 2016. A scalable tree boosting system, In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining San Francisco, CA, USA. 785–794 <https://doi.org/10.1145/2939672.2939785>

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

- Costello, J.C., Heiser, L.M., Georgii, E., Gönen, M., Menden, M.P., Wang, N.J., Bansal, M., Ammad-uddin, M., Hintsanen, P., Khan, S.A., Mpindi, J.P., Kallioniemi, O., Honkela, A., Aittokallio, T., Wennerberg, K., NCI DREAM Community, Collins, J.J., Gallahan, D., Singer, D., Saez-Rodriguez, J., Kaski, S., Gray, J.W., Stolovitzky, G. 2014. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol.* 32(12):1202-12.
<https://doi.org/10.1038/nbt.2877>.
- Fearon, E.R., Vogelstein, B.1990 A genetic model for colorectal tumorigenesis. *Cell*, **61(5)**,759-67.
[https://doi.org/10.1016/0092-8674\(90\)90186-i](https://doi.org/10.1016/0092-8674(90)90186-i).
- Gao, Y., Lyu, Q., Luo, P., Li, M., Zhou, R., Zhang, J., & Lyu, Q. 2021. Applications of Machine Learning to Predict Cisplatin Resistance in Lung Cancer. *International Journal of General Medicine*, **14**, 5911–5925.
<https://doi.org/10.2147/IJGM.S329644>
- Geeleher, P., Zhang, Z., Wang, F., Gruener, R.F., Nath, A., Morrison, G., Bhutra, S., Grossman, R.L., Huang, R.S. 2017. Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. *Genome Res*, **27(10)**, 1743-1751.
<https://doi.org/10.1101/gr.221077.117>
- Gillet, J.P., Calcagno A.M., Varma, S., Marin, o M., Green, L.J., Vora, M.I., Patel C., Orina J.N., Eliseeva, T.A., Singal, V., Padmanabhan, R., Davidson, B., Ganapathi, R., Sood A.K., Rued,a B.R., Ambudkar S.V., Gottesman, M.M. 2011. Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. *Proc Natl Acad Sci U S A*. 2011 Nov 15;108(46):18708-13.
<https://doi.org/10.1073/pnas.1111840108>
- Gönen, M. and Margolin, A.A. 2014. Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning. *Bioinformatics*, **30(17)**, 556-63.
<https://doi.org/10.1093/bioinformatics/btu464>
- Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer, New York, 150-190
- Joo, M., Park, A., Kim, K., Son, W-J., Lee, H.S., Lim, G., Lee, J., Lee, D.H., An, J., Kim, J.H. 2019 A Deep Learning Model for Cell Growth Inhibition IC50 Prediction and Its Application for Gastric Cancer Patients. *International Journal of Molecular Sciences*. **20(24)**:6276.
<https://doi.org/10.3390/ijms20246276>
- Langdon, S.P. 2004. Cell culture contamination: an overview. *Methods Mol Med*. 2004, **88**:309-17.
<https://doi.org/10.1385/1-59259-406-9:309>
- Lecun, Y., Bottou, L., Bengio, Y., P. Haffner, 1998. "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, 86, 11, 2278-2324
<https://doi.org/10.1109/5.726791>
- Masters, J.R.2000. Human cancer cell lines: fact and fantasy. *Nat Rev Mol Cell Biol.* **1(3)**, 233-6.
<https://doi.org/10.1038/35043102>
- Matlock, K., De Niz, C., Rahman, R., Ghosh, S., Pal, R. 2018. Investigation of model stacking for drug sensitivity prediction. *BMC Bioinformatics*, **19(3)**, 71.
<https://doi.org/10.1186/s12859-018-2060-2>
- Park, A., Joo, M., Kim, K., Son, W.J., Lim, G., Lee, J., Kim, J.H., Lee, D.H., Nam, S. 2022. A comprehensive evaluation of regression-based drug responsiveness prediction models, using cell viability inhibitory concentrations (IC50 values). *Bioinformatics*, **38(10)**, 2810-2817.
<https://doi.org/10.1093/bioinformatics/btac177>
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A., 2018. CatBoost: unbiased boosting with categorical features, Advances in neural information processing systems, 31
- Roden, D.M., McLeod, H.L., Relling, M.V., Williams, M.S., Mensah, G.A., Peterson, J.F., Van Driest, S.L., 2019 *Pharmacogenomics. Lancet.*, **394(10197)**:521-532.
[https://doi.org/10.1016/S0140-6736\(19\)31276-0](https://doi.org/10.1016/S0140-6736(19)31276-0)
- Sharma, A. and Rani, R. 2020. Ensembled machine learning framework for drug sensitivity prediction. *IET Syst Biol*, **14(1)**, 39-46.
<https://doi.org/10.1049/iet-syb.2018.5094>
- Singh, D. P. and Kaushik, B. 2023. A systematic literature review for the prediction of anticancer drug response using various machine-learning and deep-learning techniques. *Chemical Biology & Drug Design*, **101(1)**, 175-194.
- Steele, F.R. 2009. Personalized medicine: something old, something new. *Per Med.*, **6(1)**:1-5.
<https://doi.org/10.2217/17410541.6.1.1>
- Vis, D.J., Bombardelli, L., Lightfoot, H., Iorio, F., Garnett, M.J., Wessels, L.F. 2016. Multilevel models improve precision and speed of IC50 estimates. *Pharmacogenomics*. 2016 May, **17(7)**, 691-700.
<https://doi.org/10.2217/pgs.16.15>
- Wang, H., Xi, J., Wang, M., Li, A. 2020. Dual-Layer Strengthened Collaborative Topic Regression Modeling for Predicting Drug Sensitivity. *IEEE/ACM Trans Comput Biol Bioinform.*, 2020 Mar-Apr, **17(2)**:587-598.
<https://doi.org/10.1109/TCBB.2018.2864739>
- Wei, D., Liu, C., Zheng, X., Li, Y. 2019. Comprehensive anticancer drug response prediction based on a simple cell line-drug complex network model. *BMC Bioinformatics*. 2019 Jan 22, **20(1)**:44.

<https://doi.org/10.1186/s12859-019-2608-9>

van der Meer, D., Barthorpe S., Yang W., Lightfoot H., Hall, C., Gilbert, J., Francies, H.E., Garnett, M.J., 2019, Cell Model Passports-a hub for clinical, genetic and functional datasets of preclinical cancer models. *Nucleic Acids Res.*, Jan, **8;47**(D1):D923-D929. <https://doi.org/10.1093/nar/gky872>.

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

Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., Ramaswamy, S., Futreal, P.A., Haber, D.A., Stratton, M.R., Benes, C., McDermott, U., Garnett, M.J. 2013. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, 2013 Jan, **41** (Database issue):D955-61. <https://doi.org/10.1093/nar/gks1111>

İnternet Kaynakları

[1] <https://cellmodelpassports.sanger.ac.uk/downloads>

[2] <https://www.cancerrxgene.org/celllines>

[3] <https://cellmodelpassports.sanger.ac.uk>