TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TOPLULUK YÖNTEMİ VE İLAÇ İMZALARI KULLANILARAK ANTİ KANSER İLAÇLARIN AKTİVİTE TAHMİNİ

YÜKSEK LİSANS TEZİ

Ertan TOLAN

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Yrd. Doç. Dr. Mehmet TAN

ARALIK 2016

Fen Bilimleri Enstitüsü Onayı

....
Prof. Dr. Osman EROĞUL
Müdür

Bu tezin Yüksek Lisans derecesinin tüm gereksinimlerini sağladığını onaylarım.

....
Doç. Dr. Oğuz ERGİN
Anabilimdalı Başkan Vekili

TOBB ETÜ, Fen Bilimleri Enstitüsü'nün 141111036 numaralı Yüksek Lisans öğrencisi **Ertan TOLAN** 'ın ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"TOPLULUK YÖNTEMİ VE İLAÇ İMZALARI KULLANILARAK ANTİ KANSER İLAÇLARIN AKTİVİTE TAHMİNİ"** başlıklı tezi 19.12.2016 tarihinde aşağıda imzaları olan jüri tarafından kabul edilmiştir.

Tez Danışmanı: **Yrd. Doç. Dr. Mehmet TAN** ....
TOBB Ekonomi ve Teknoloji Üniversitesi

Jüri Üyeleri: **Doç. Dr. Pınar KARAGÖZ (Başkan)** ....
Orta Doğu Teknik Üniversitesi

Doç. Dr. Osman ABUL ....
TOBB Ekonomi ve Teknoloji Üniversitesi

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, alıntı yapılan kaynaklara eksiksiz atıf yapıldığını, referansların tam olarak belirtildiğini ve ayrıca bu tezin TOBB ETÜ Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırlandığını bildiririm.

Ertan TOLAN

ÖZET

Yüksek Lisans Tezi

TOPLULUK YÖNTEMİ VE İLAÇ İMZALARI KULLANILARAK ANTİ KANSER İLAÇLARIN AKTİVİTE TAHMİNİ

Ertan TOLAN

TOBB Ekonomi ve Teknoloji Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Yrd. Doç. Dr. Mehmet TAN

Tarih: ARALIK 2016

Kişiselleştirilmiş kanser tedavisi, kanserin karmaşıklığı da göz önünde bulundurulduğunda, gelişmekte olan bir yaklaşımdır. Kişiselleştirilmiş tedavinin bir parçası olarak, bir ilacın bir hücre hattındaki etkinliği labarotuvar ortamında ölçülür. Ancak, bu deneylerin yapılması çok zordur ve önemli bir maddi kaynak gerektirir. Bu zorlukların üstesinden gelmek için hesaplayıcı yöntemler, sağlanan veri kümeleri ile birlikte bilgisayar ortamında kullanılır.

Bu çalışmada aktivite tahmini problemi bir regresyon problemi olarak ele alınmıştır ve öncelikle her bir ilaç-hücre hattı çiftinin tahmin hatasının azaltılması için üç farklı regresyon modeli birleştirilerek bir topluluk modeli tasarlanmıştır. Temel modeller; gradyan destekli regresyon, çekirdekli bayes çoklu-iş öğrenme ve iz-norm regülarizasyonlu çoklu-iş öğrenme olarak tanımlanmıştır. Oluşturulan modeli değerlendirmek için iki büyük veri kümesi, genomics of drug sensitivity in cancer ve cancer therapeutics response portal, kullanılmıştır. Bu değerlendirmenin sonuçları, topluluk yönteminin tahminlerinin temel modellerin her birinin tek başlarına yaptıkları tahminlerden önemli ölçüde daha iyi olduğunu göstermektedir. Bunun sonucunda orijinal veri kümelerinde görülmeyen ilaç-hücre hattı çiftleri için oluşturulan modelin sitotoksisite tahminleri rapor edilmiştir. Araştırmacılar tarafından yapılan canlı içi (in vivo) laboratuvar çalışmaları da, rapor sonuçlarını desteklemektedir.

İlaç aktivitelerini tahmin etmesi için oluşturulan bir diğer model de imza benzerlik tabanlı regülarizasyonlu çoklu-iş öğrenme modelidir. Bu modelin oluşturulması için öncelikle LINCS veri kümesinde bulunan ilaç-hücre hattı deneyleri incelenerek ilaçlar için aktivite imzaları oluşturulmuştur. Oluşturulan bu ilaç imzaları ile ilaçların arasındaki benzerlikler hesaplanarak, ilaçların benzerliklerini gözardı eden modellerden daha güçlü bir tahmin edici model üretilmiştir. Bu model yine genomics of drug sensitivity in cancer ve cancer therapeutics response portal veri kümeleri kullanılarak değerlendirilmiştir ve her iki veri kümesinde de karşılaştırılan modellere göre belirgin bir üstünlük sağlandığı gözlenmiştir.

Anahtar Kelimeler: Kanser, Aktivite tahmini, Regresyon, Yapay öğrenme, Çoklu-iş öğrenme, Topluluk modeli, İlaç aktivite imzası

ABSTRACT

Master of Science

ACTIVITY PREDICTION OF ANTI CANCER DRUGS BY USING ENSEMBLE LEARNING AND DRUGS' SIGNATURES

Ertan TOLAN

TOBB University of Economics and Technology
Institute of Natural and Applied Sciences
Department of Computer Engineering

Supervisor: Asst. Prof. Mehmet TAN

Date: DECEMBER 2016

Personalized cancer treatment is an ever-evolving approach due to complexity of cancer. As a part of personalized therapy, effectiveness of a drug on a cell line is measured at the laboratory environments. However, these experiments are backbreaking and money consuming. To surmount these difficulties, computational methods are used with the provided data sets.

In the present study, we considered this as a regression problem and firstly designed an ensemble model by combining three different regression models to reduce prediction error for each drug-cell line pair. We defined our base models as gradient boosting regression, kernelized bayesian multi-task learning and trace-norm regularized multi-task learning. Two major data sets, genomics of drug sensitivity in cancer and cancer therapeutics response portal, were used to evaluate our method. Results of this evaluation show that predictions of ensemble method are significantly better than models *per se*.

Furthermore, we report the cytotoxicty predictions of our model for the drug-cell line pairs that do not appear in the original data sets.

The another method to predict anti cancer drug activity is similarity of signature regularized multi-task learning. To constitute this model, firstly, drug signature is generated by examining drug-cell line experiments found in LINCS data set. Then, by using these activity signatures of drugs, similarities between drugs are calculated and a powerful model which overperforms the models which ignore drug similarities is designed.

Also this model is evaluated with genomics drug sensitivity in cancer and cancer therapeutics response portal data sets and results of the both data sets show that constituted model have significantly more predictive power than contrasted model.

Keywords: Cancer, Activity prediction, Regression, Machine learning, Multi-task learning, Ensemble learning, Drug activity signature

TEŞEKKÜR

Yüksek lisans eğitimim ve tez çalışmalarım boyunca beni destekleyen değerli hocam Yrd. Doç. Dr. Mehmet TAN ’a,

Öğrenim hayatım boyunca sağladığı burs imkanı ile ve de sunduğu çalışma ortamıyla beni destekleyen TOBB Ekonomi ve Teknoloji Üniversitesi ailesine,

Tecrübelerinden faydalandığım TOBB Ekonomi ve Teknoloji Üniversitesi bilgisayar mühendisliği bölümünün değerli öğretim üyelerine,

Birlikte çalışmaktan mutluluk duyduğum yüksek lisans arkadaşlarıma sonsuz teşekkürlerimi sunarım.

Bu tez kapsamında yapılan çalışmalar TÜBİTAK (Proje No: 115E274) tarafından desteklenmektedir.

İÇİNDEKİLER

ŞEKİL LİSTESİ

ÇİZELGE LİSTESİ

Sayfa

KISALTMALAR

CTRP : Kanser Tedavi Tepki Portalı
GBR : Gradyan Destekli Regresyon
GDSC : Kanserde İlaç Hassasiyet Genomiği
İBTRÇÖ : İmza Benzerliği Tabanlı Regülarizasyonlu Çoklu-İş Öğrenme
WAMSE : İlaçların Ortalama Karesel Hatalarının Ortalaması
AMSE : İlaçların Ortalama Karesel Hatalarının Ağırlıklı Ortalaması
LINCS : Tümleşik Ağ Tabanlı Hücresel İmza Kütüphanesi
KBMTL : Çekirdekli Bayes Çoklu-İş Öğrenme
RBF : Radial Basis Function
SRMTL : Seyrek Yapı Regülarizasyonlu Çoklu-İş Öğrenme
NDPB : Tahmin Edicinin En İyi Olarak Tahmin Ettiği İlaç Sayısı
TRMTL : İz-norm Regülarizasyonlu Çoklu-İş Öğrenme

SEMBOL LİSTESİ

Bu çalışmada kullanılmış olan simgeler açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	**Açıklama**
AUC	Doz-tepki eğrisi altındaki alan
CL	Hücre hattı
dr	İlaç
$gen\uparrow$	Yukarı yönlü regüle edilmiş gen
$gen\downarrow$	Aşağı yönlü regüle edilmiş gen
IC_{50}	Yarı maksimum durdurucu konsantrasyon değeri
MSE	Mean Squared Error
ρ_1	Benzerlik tabanlı regülarizasyon paremetresi
ρ_2	ℓ_1 norm regülarizasyon paremetresi
ρ_{L2}	ℓ_2 norm regülarizasyon paremetresi

1. GİRİŞ

Kanser, hücrenin genetik yapısındaki bozulmalar sonucu vücudun çeşitli bölgelerindeki kontrol dışı çoğalma ile oluşan hastalıkların genel adıdır. Hastalığa yakalanma ve ölüm oranı bakımından kanser, dünya çapında önde gelen hastalıklardan biri olarak düşünülmektedir. 2016 yılında ABD'de yaklaşık 1.685.210 yeni kanser vakası teşhisi konacağı ve 595.690 kişinin hastalık dolayısıyla öleceği tahmin edilmektedir [34]. Kanser teşhisinde ve tedavisinde; kanserin ilerlemesine ve ilk gözlemlendiği bölgeye göre çok farklı yöntemler kullanılabilir. Bu tedavi ve teşhis yöntemlerinin başlıcaları cerrahi, kemoterapi, ışın tedavisi, immünoterapi olarak sayılabilir. Bu yöntemler tek başlarına kullanılabildiği gibi birlikte de kullanılabilirler. Örneğin; kemoterapi ve ışın tedavisi çoğu zaman birlikte kullanılan yöntemlerdir.

Bilinen yüzden fazla kanser çeşidi vardır ve genellikle kanserin şekillenmeye başladığı yere göre isim alır. Örneğin, deride ve dokuda başlayıp iç organlara yayılan kanser karsinoma olarak adlandırılırken; kemik, kıkırdak, yağ, kas veya kan damarları gibi bağ veya destek dokularında başlayan kanser sarkoma olarak adlandırılır. Kemik iliği gibi kan yapan dokularda başlayıp çok sayıda anormal kan hücresinin üretilmesine neden olan kanser de lösemi adını alır. Bunun yanında vücutta başladığı yere göre de isim alabilmektedir. Örneğin meme kanseri; meme dokusunda, akciğer kanseri; akciğer dokusunda başlayan ve kontrol edilemeyen hücre çoğalmasıdır.

Anti kanser ilaçlar kullanarak bu çeşitli kanser hastalıklarını tedavi etmeye çalışan yöntem olarak bilinen kemoterapi, en çok tercih edilen tedavi yöntemidir [23]. Geleneksel kemoterapide, farklı insanlarda oluşan ve birbirine benzediği düşünülen kanser hastalıklarının tedavisinde benzer ilaçlar kullanılmıştır. Ancak daha sonra bu kanserli hücrelerin genetik olarak birbirine benzemediği yapılan araştırmalarda ortaya konulmuştur [2, 10]. Buradan hareketle yapılan çalışmalarda da aynı tip kanser hastalarına aynı tedavilerin uygulanmasının aslında doğru olmadığı gösterilmiştir [10]. Kontrol dışı çoğalan kanserli hücreleri öldürmeyi amaçlayan anti kanser ilaçları, aynı zamanda sağlıklı hücrelere de zarar verebilmektedir. Kemoterapi tedavisinin hasta üzerinde bıraktığı yan etkiler de düşünüldüğünde, tedavi için doğru ilacın bulunması önem arz etmektedir. Dolayısıyla araştırmacılar, tedavide kullanılacak ilacın çeşidine, dozuna ve ne şekilde kullanılacağına karar vermek için labarotuvar çalışmalarına yoğunlaşmışlardır.

Geleneksel tedavi yöntemleri yerine, kişiselleştirilmiş kanser tedavisi ile, kanser teşhisi konulan hastadan alınan tümör hücresinin yapısı laboratuvar ortamında incelenerek hastanın tedavisine yön verilir. İlgili hasta için, doğru zamanda doğru tedavi yöntemini bulmayı amaçlayan kişiselleştirilmiş kanser tedavisinin, geleneksel terapilere göre daha etkili yöntem olduğu artık bilinen bir gerçektir [19].

Kişiselleştirilmiş tedavinin bir parçası olarak, kanser teşhisi konulmuş bir hastadan alınan tümör hücreleri üzerindeki deneyler, bir tümör hücresinin bir kanser ilacı için ne kadar hassas olduğunu gösterir. Belirli bir ilacın belirli bir kanser tedavisinde etkili olup olmayacağına karar verilmesinde, genellikle kanser hücre hattı üzerindeki bu deneysel sonuçlar başlangıç noktası olur. Ancak, potansiyel ilaç adaylarının sayıca fazla olması nedeniyle, geniş kapsamlı kimyasal bileşik-hücre hattı çiftlerinin deneyleri önemli bir maliyet oluşturmaktadır. Bu sorunun üstesinden gelmek için, tümör hücrelerinin uygulanan ilaçlara tepkisini, laboratuvarda deneyler yaparak ölçmek yerine, ilaç tepkilerini tahmin etmek için kullanılan hesaplama modelleri tasarlanmıştır.

Yapay öğrenme modelleri, son zamanlarda büyük ölçekli ilaç tepki veri tabanlarının [1, 28, 30, 33, 39] yayımlanması sayesinde ilaç aktivite tahmininde kullanılır hale gelmiştir. Bu veri tabanları, yüzlerce kanser hücre hattına karşı çok sayıda kimyasal bileşiğin sitotoksisite deneylerinin sonuçlarını içermektedir. Hücre sitotoksisite değerleri, ilaçların hücreleri yok etme kabiliyetinin ölçülmesinde kullanılır. Veri tabanlarında bu veriler önişlemeden geçirilerek hesaplama modelleri için anlamlı hale getirilir. Örneğin sitotoksisite değeri için GDSC (Genomics of Drug Sensitivity in Cancer) veri kümesinde, doz-tepki eğrisinin altında kalan alana ve yarı maksimum durdurucu konsantrasyon değerinin doğal logaritmasına yer verilir.

Anti kanser ilacının tümör hücresi üzerindeki etkinliği hakkında bilgi veren doz-tepki eğrisi altındaki alan (AUC) ve yarı maksimum durdurucu konsantrasyon (IC_{50}) değerlerinin tahmin edilmesi için oluşturulan modeller açısından bu problem sınıflandırma ya da regresyon problemi olarak değerlendirilebilir. Sınıflandırmada, bir hücrenin ilaca hassas olup olmadığı öngörülmeye çalışılırken, regresyonda, IC_{50} veya AUC cinsinden sitotoksisitenin tam değerinin tahmin edilmesi amaçlanır. İkili sınıflandırma (hücre ilaca hassastır veya dirençlidir) ile karşılaştırıldığında, regresyon (ilacın sitotoksisite değerinin tam olarak tahmini) açıkça daha zordur, ancak bununla beraber ilacın tümör hücresini nasıl etkilediği konusunda çok daha fazla bilgi verir.

Buradaki en önemli hususlardan biri, hücre hatlarının nasıl karakterize edileceğidir. Hücre hatları; gen ifadesi, DNA metilasyonu ve kopya sayısı varyasyonu verileri gibi birkaç farklı veri türü kullanılarak karakterize edilebilir. Bunların arasında; gen ifadesi verileri, en bilgilendirici olarak görülür ve son zamanlarda yapılan 'DREAM challenge' da bunu doğrulamaktadır [8]. Gen ifadesi profili, hücresel fonksiyonların genel bir görüntüsünü oluşturmak için binlerce genin aktivitesinin ölçülmesi ile elde edilir. Bu profiller, aktif olarak bölünen hücreleri ayırt edebilir veya hücrelerin belirli bir tedaviye nasıl tepki gösterdiğini ölçebilir. Bu sebeple çalışmalarımızda hücre hatları, gen ifadesi profilleri kullanılarak gösterimlenmiştir.

Topluluk öğrenme, farklı öğrenme algoritmalarının çeşitli yollarla birleştirilerek daha genel bir modelin oluşturulduğu öğrenme yöntemidir. Topluluk yöntemlerinde, temel öğrenme algoritmalarının her birinden elde edilebilenden daha iyi tahmin edici performans elde etmek için, çoklu öğrenme algoritmaları kullanılır [11].

Bu tez çalışmasının ilk bölümünde ilaç tepkilerinin tam değerini tahmin etmek için üç farklı yöntemi bir araya getiren bir topluluk modeli önerilmektedir. Temel olarak alınan üç model; gradyan destekli regresyon, çekirdekli bayes çoklu-iş öğrenme ve iz-norm regülarizasyonlu çoklu-iş öğrenmedir. Çoklu-iş öğrenmenin amacı, birden fazla iş için ortaklaşa öğrenerek öğrenme algoritmalarının performansını arttırmaktır. Anti kanser ilaçlar çoklu-iş öğrenme modellerindeki birbirleriyle ilişkili işler olarak ele alınabildiği için [40], ortak öğrenme kanser ilaçlarının aktivite tahminine uygun bir modeldir ve bu alandaki uygulanabilirliği çok yüksektir. Bu sebeple topluluk modeli için iki tanınmış ve açık kaynak olarak paylaşılmış çoklu-iş öğrenme modeli seçilmiştir; KBMTL (Kernelized Bayesian Multi-task Learning) (Çekirdekli Bayes Çoklu-İş Öğrenme) ve TRMTL (Trace-norm Regularized Multi-task Learning) (İz-norm Regülarizasyonlu Çoklu-İş Öğrenme). Ayrıca GBR (Gradient Boosting Regression) (Gradyan Destekli Regresyon), tahmin gücü dikkate alınarak tek-iş öğrenme modelleri arasından seçilmiştir. Bu üç yöntem birleştirilerek her bir ilacın tahmin hata oranının ve tüm ilaçların ağırlıklı hata ortalamasının düşürülmesi amaçlanmıştır. Tahmin modeli oluşturmak için, GDSC (Genomics of Drug Sensitivity in Cancer) [39] ve CTRP (Cancer Therapeutics Response Portal) [28, 30] veri setleri tarafından sağlanan yüzlerce hücre hattı ve ilaç tepki verisi kullanılmaktadır.

Tez çalışmasının ikinci kısmında, SRMTL (seyrek çizge regülarizasyonlu çoklu-iş öğrenme [43, 44] algoritması kullanılarak ilk kısımda belirtilen problemi ilaç benzerlik ilişkilerini ele alarak çözmeye çalışan başka bir model oluşturulmuştur. SRMTL, parametre olarak verilen çoklu işlerin benzerliği çizgesini kullanarak regülarizasyon işlemini yapan bir modeldir. Bu benzerlik ilişkilerinin modele parametre olarak verilmesi, çoklu-iş öğrenme algoritmasının hedeflediği ’işlerin birlikte öğrenilmesi’ görevinin daha iyi bir şekilde yapılmasını amaçlamaktadır. Bu modelin ilaç aktivite tahmini için uygun olduğu düşünülerek imza benzerlik tabanlı regülarizasyonlu çoklu-iş öğrenme (İBTRÇÖ) algoritması, ilaçların hücre hatları üzerindeki hassasiyetinin tahmini için geliştirilmiştir. Gerekli olan ilaç benzerliklerinin hesaplanması için öncelikle her bir ilaç için, ilaçların hücre hatları üzerindeki aktiviteleri göz önünde bulundurularak aktivite imzası oluşturulmuştur. Bu işlem için LINCS (The Library of Network-Based Cellular Signatures) [13] veri kümesinden yararlanılmıştır. Bu veri kümesi aracılığıyla, ilaçların hücre hatları ile girdikleri tepkime sonrası gen ifadelerini nasıl değiştirdikleri bilgisine ulaşılabilmektedir. Tepkime sonrası genlerin aşağı ya da yukarı yönlü regüle olduğu bilgisi kullanılarak her ilaç için bir aktivite imzası tasarlanmıştır. Dolayısıyla benzerlik ilişkisi, aynı genleri aynı yönlerde regüle eden iki ilacın birbirine benzediği hipotezine dayandırılarak hesaplanmıştır.

İlaçların benzerlik ilişkisini kullanan bu model, GDSC ve CTRP veri kümeleri üzerinde, bu veri kümelerinde bulunan ilaçların aktivite imzaları ayrı ayrı oluşturularak değerlendirilmiştir. Bu değerlendirme sonucunda her iki veri kümesi için de modelin, benzerlik ilişkisinin verilmediği modellerden daha başarılı sonuçlar verdiği gözlenmiştir.

Bu tez çalışması şu şekilde organize edilmiştir. Bölüm 1’de tez çalışmasında ele alınan problem tanımlanmış ve problemin ortaya çıkışına ve çözümüne yönelik genel bilgiler verilmiştir. Bölüm 2’de problem hakkında yapılan benzer çalışmalar gruplandırılarak özetlenmiştir. Problemin çözümüne yönelik, tez çalışmasında kullanılan yöntemler ve oluşturulan topluluk modeli detaylı olarak Bölüm 3’de ele alınmıştır. Bölüm 4’de imza oluşturma ve benzerlik hesaplama adımları detaylandırılarak, imza benzerlik tabanlı regülarizasyonlu çoklu-iş öğrenme modeli anlatılmıştır. Bölüm 5’de tez çalışmasında kullanılan veri kümeleri ayrıntılı olarak anlatılmıştır. Oluşturulan topluluk modelinin sonuçları temel modeller ile; benzerlik tabanlı regülarizasyonlu modelin sonuçları benzerlik bilgisinin verilmediği model ile karşılaştırmalı olarak verilmiştir. Tez çalışmasının sonuç kısmı ise Bölüm 6'de bildirilmiştir. Bu bölümde ayrıca oluşturulan modellerin deneysel sonuçları incelenerek, potansiyel gelecek çalışmalara yer verilmiştir.

2. İLGİLİ ÇALIŞMALAR

Bu bölümde anti kanser ilaçları için aktivite tahmini yapan literatürdeki çalışmalar incelenecektir. Çalışmalar genel olarak kullandıkları veri türleri ve modelleme teknikleri açısından farklılık gösterir.

2.1 Kullanılan Veri Türlerine Göre Çalışmalar

Farklı veri türü olarak; tekli nükleotid mutasyonlar, gen kopyalama sayıları ve gen ifadesinden oluşan herkese açık ya da özel veri kümeleri ile birçok model geliştirilmiştir [8, 20]. Tek veya çoklu kaynaklara dayanan farklı model örnekleri, örneğin gen ifadesi verilerinin gen kopya sayısı verileri ile kombinasyonu, yalnızca gen ifadesi verileri ile en yaygın kullanılan kaynak olarak bildirilmiştir. Karşılaştırmalı analizler genellikle gen ifadesi verisinin en güçlü öngörme özelliklerini içerdiğini ve bu entegre modellerin ilaç tepki tahminlerinin doğruluğunu dikkate değer bir şekilde artırabildiğini göstermiştir [8].

Wan [36] çalışmasında Cancer Cell Line Encyclopedia (CCLE) için iki, DREAM-Challenges için beş farklı genomik karakterizasyon veri kümesi kullanarak heterojen bir yapı oluşturmuştur. CCLE veri kümesi için; gen ifadesi verileri ve mutasyon bilgilerini içeren SNP6 verileri kullanılmıştır. Bu iki veri kümesinden, ilaçların insan kinazlarını hedef aldığını göz önünde bulundurarak, yalnızca kinaz üretilen 400 gen öznitelik olarak seçilmiştir. Rastgele orman yöntemiyle CCLE veri kümesi üzerinde yaptığı değerlendirmede, çoklu genomik karakterizasyonun hata oranını düşürdüğünü belirtmiştir. DREAM-Challenges verileri için ise 5 farklı veri kümesinden oluşturulan bütün kombinasyonlardan birer sonuç üreterek, bu sonuçları topluluk yöntemiyle birleştirmiştir. Burada veri kümesi sayısının arttırılması, deney maliyetini de arttırmaktadır. Ayrıca ileride verilerin artacağı da düşünüldüğünde tüm veri kümelerinin kombinasyonundan sonuç alarak bunları topluluk modeli olarak kullanmak çalışma zamanı açısından da problem oluşturabilir.

Genelde öznitelik olarak hücre hatlarının genomik bigileri kullanılırken, hücre hatlarının genomik karakterizasyonunu ilaçların yapısal özellikleri ile birleştirerek oluşturulmuş modeller de vardır[22] . Ek olarak, ilaçların kimyasal bilgilerininin ve hücre hatlarının profil verisinin birleşiminin girdi olarak alınarak ilaç duyarlılığının tahmininin geliştirilebileceği, NCI-60 verileri üzerinde ilaçlar için modeller eğitilerek gösterilmiştir[7] . Başka bir çalışmada da araştırmacılar modelini hücre hatlarının gen ifadesi seviyelerinin çok değişkenli etkileşimiyle ilişkilendirerek ilaçlar için aktivite tahminini güçlendirmiştir[29] .Bu çalışmalarda genel olarak ileride veri sayısı ve çeşidinin artmaya devam edeceği göz önünde bulundurulmuştur. Oluşturulacak olan modelde tek bir veri yerine birden fazla veri kullanılmasının tahmin gücünü iyileştireceği belirtilmektedir. Bu modellerin ileride, veri miktarı ve çeşidi arttıkça, daha iyi performans göstereceklerı düşünülebilir.

2.2 Modelleme Yöntemlerine Göre Çalışmalar

İlaç aktivite tahmini için tasarlanan modeller daha çok gözeticili öğrenme tekniklerine dayanmaktadır. Ancak gözeticisiz tekniklerden de, gözeticili tahmin modellerinin oluşturulmasında yararlanılabilmektedir [4, 17, 25]. Ayrıca çalışmalar çoklu-iş öğrenme ve tekli-iş öğrenme modelleri olarak da incelenebilir. Çoklu-iş öğrenme modeli, benzer ilaçların öğrenme aşamasında birbirinden yararlanmasını amaçlarken, tekli-iş öğrenmede modeller ayrık olarak öğrenirler.

Gözeticili teknikler genel olarak regresyon ve sınıflandırma modelleri olarak iki ana başlık altında incelenebilir. Regresyon modelleri, hesaplama tekniklerini kullanır ve ilaç hassasiyetini IC_{50} veya AUC değerleri cinsinden tahmin etmeye çalışır [14, 24]. Sınıflandırma modellerinde ise önceden belirlenen bir değeri göz önünde bulundurarak ilacın hücre hattı üzerindeki aktivitesi hassas ya da dirençli olarak tahmin edilmeye çalışılır [15, 20].

Anti kanser ilaç aktivite tahmini için çoklu-iş ya da tek-iş öğrenme algoritmaları kullanılmıştır. Çoklu-iş öğrenme modelinin ilaç aktivite tahmini için uygun bir model olduğu gösterilmiştir [40]. Bu çalışmada regülarizasyon için kullanılan iz-norm problemi çarpanları dönüşümlü yönlendirme yöntemi (ADMM) ile çözülerek model üretilmiştir. Üretilen bu model farklı sayıdaki ilaçlardan oluşan kümeler ile eğitilmiştir. Alınan sonuçlar üç farklı veri kümesi kullanılarak elastik net ile karşılaştırıldığında daha iyi bir performans gösterdiği gözlenmiştir.

Tan [35] çalışmasında ilaçlar arasındaki ilişkilerin doğrusal olmadığı hipotezini öne sürmüş ve bu ilişkileri ortaya çıkarmak için modelinde doğrusal olmayan çekirdek kullanmıştır. Bu çalışmada oluşturulan çekirdekli model, iz-norm ile kullanılarak benzer yapıları hedefleyen ilaçlar arasındaki ilişkilerden çoklu-iş öğrenme yöntemiyle faydalanmıştır.

Veri tabanlarındaki hücre hatları ve ilaçlar için tüm deneyler yapılamamıştır ve bu sebeple bazı kayıp değerler bulunmaktadır. Bu kayıp değerli örnekler çıkarıldığında ise veri tabanları küçülmektedir. Modeller bu kayıp verileri kullanıp kullanmamasına göre de farklılaşabilir. Gönen [18], meydana gelen bu küçülmenin önüne geçmek için kayıp değerli örneklerin de kullanıldığı çekirdekli bayes çoklu-iş öğrenme modelini tasarlamıştır. Ayrıca kullandığı çekirdek ile deneysel gürültüleri de azaltmıştır. Gönen [18], çalışmaları sonucunda öğrenme modelinin anti kanser ilaçlarının tepkisini tahmin etmek için oldukça kullanılabilir olduğunu göstermiştir ve çalışmalarını açık olarak paylaşmaktadır. Bu modelin fazla sayıdaki parametre listesi uygulanabilirliğini azalttığı söylenebilir ancak yöntem sahibi tarafından sağlanan parametreler ile birlikte bu tez çalışmasında da kullanılmıştır.

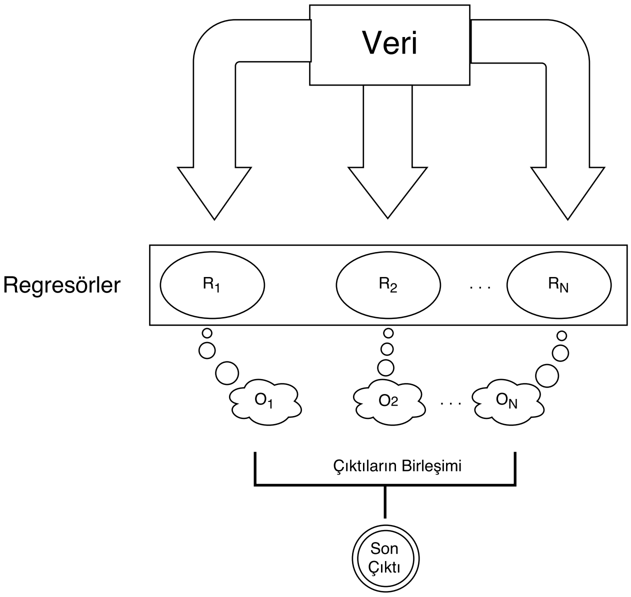
Bunların yanı sıra farklı modelleme teknikleri de kullanılmıştır. İlaç aktivite tahmini için, benzer hücre hatlarının ve benzer ilaçların benzer tepkiler vereceği hipotezini temel alarak hücre hattı benzerlik ağı ve ilaç benzerlik ağı oluşturulmuştur [41].

Hücre hattı benzerlik ağını, hücre hatlarının gen ifadesi profillerinin ikili Pearson korelasyonu ile hesaplarken, ilaç benzerlik ağı için ise 1-D ve 2-D ilaç yapılarının ikili Pearson korelasyonu kullanılmıştır. Bu benzerlik ağları doğrusal ağırlıklı bir model ile bütünleştirilerek, tek katmanlı modellerden daha iyi performans gösteren entegre bir ağ önerilmiştir. Farklı ilaçların sinerjik etkilerinden yararlanarak etkili kanser tedavileri geliştirmenin bir başka yöntemi [26] tarafından sunulmuştur. Çalışmalar genel olarak veri tabanları üzerinde ayrı ayrı yapılırken, farklı olarak, Dong [12], kendi modelini oluşturmak ve değerlendirmek için iki farklı veri tabanını, CCLE ve GDSC'yi birleştirmiştir.

Öznitelik seçimi veriyi ilgisiz özniteliklerden arındırarak verinin boyutunu azaltan bir işlemdir. Çoklu örnek öğrenimi (multiple instance learning) ise tek tek örnekler almak yerine, öğreniciye her biri birçok örnek içeren etiketlenmiş bir dizi örnek verir. Zhao, Modelini öznitelik seçimi ve çoklu örnek öğrenimi kullanarak geliştirmiştir[42] .

3. TOPLULUK YÖNTEMİNİ KULLANARAK AKTİVİTE TAHMİNİ

Bu bölümde öncelikle kullanılan temel yapay öğrenme yöntemleri ve modellerin oluşturulması aşamasında kullanılan diğer yöntemler ve bunların nasıl seçildikleri, daha sonra ise temel yapay öğrenme algoritmaları birleştirilerek oluşturulan topluluk modeli(Şekil 3.1) açıklanmıştır.



Şekil 3.1: Topluluk öğrenme modelinin gösterimi. Topluluk öğrenme, farklı öğrenme algoritmalarının çeşitli yollarla birleştirilerek daha genel bir modelin oluşturulduğu öğrenme yöntemidir.

3.1 Temel Alınan Modeller

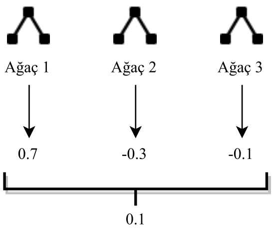
Topluluk modeli oluşturmak için öncelikle üç adet temel model seçilmiştir. Bu modellerin ilki tekli-iş öğrenme (Şekil 3.3) algoritmalarından gradyan destekli regresyondur. Diğer ikisi ise iz-norm regülarizasyonlu ve çekirdekli bayes çoklu-iş öğrenme (Şekil 3.4) algoritmalarıdır.

İz-norm regülarizasyonlu çoklu-iş öğrenme yönteminin son zamanlarda kullanıldığı çalışmaların [35, 40] sonuçlarına bakıldığında, bu yöntemde kullanılan regülarizasyonun çoklu-iş öğrenme modelleri için yarar sağladığı görülmüştür. Çekirdekli bayes çoklu-iş öğrenme ise, anti kanser ilaç aktivite tahmini için Gönen [18] tarafından kullanılmıştır ve modelin ortaya koyduğu yenilikler ile iyi bir tahmin edici olduğu gösterilmiştir. Açık kaynak olarak paylaşılan bu iki çoklu-iş öğrenme modeli temel alınan modellerden olmuştur. Gradyan destekli regresyon ise bu iki çoklu-iş öğrenme modelinin yanında, tekli-iş öğrenme modeli olarak, kendi içinde kullandığı topluluk yönteminden elde ettiği tahmin gücü de göz önünde bulundurulduğunda uygun görülmüştür. Bu yöntemlerin dışında regresyon destek vektör makinesi (SVM regression), gauss süreci regresyon (gaussian process regression), lasso regülarizasyonlu çoklu-iş öğrenme gibi yöntemler de denenmiştir. Ancak yapılan çalışmalarda bu üç model uygun görülmüştür.

3.1.1 Gradyan destekli regresyon

Regresyon, bir ya da daha fazla bağımsız değişken ile bir bağımlı değişken arasındaki ilişkiyi hesaplamaya yaran analiz işlemidir. Bu hesaplamada doğrusal modeller kullanılabileceği gibi karar ağaçları yardımıyla doğrusal olmayan modeller de üretilebilir. Oluşturulan karar ağaçlarının yapraklarında sürekli değerler yer alır ve iç düğümlerinde ise örnek girdi için verilen öznitelikler ifade edilir. Yapay öğrenmede tek bir karar ağacıyla model oluşturulabildiği gibi çok sayıda karar ağacı birlikte kullanılarak topluluk modelleri de oluşturulabilir. Örneğin destek (boosting) algoritmaları zayıf öğrenicilerden güçlü bir öğrenici elde etmek için karar ağaçlarını birlikte kullanır (Şekil 3.2).

Gradyan destekli regresyon[16] da çoğunlukla karar ağaçlarıyla temsil edilen ve birçok zayıf öğrenicinin bir araya gelmesiyle oluşan bir topluluk modelidir. Zayıf öğreniciler mevcut modele birbiri arkasından eklenerek mevcut öğrenicinin eksiklikleri telafi edilir ve zayıf öğrenicilerin birleşimiyle güçlü bir öğrenici meydana getiriliir.



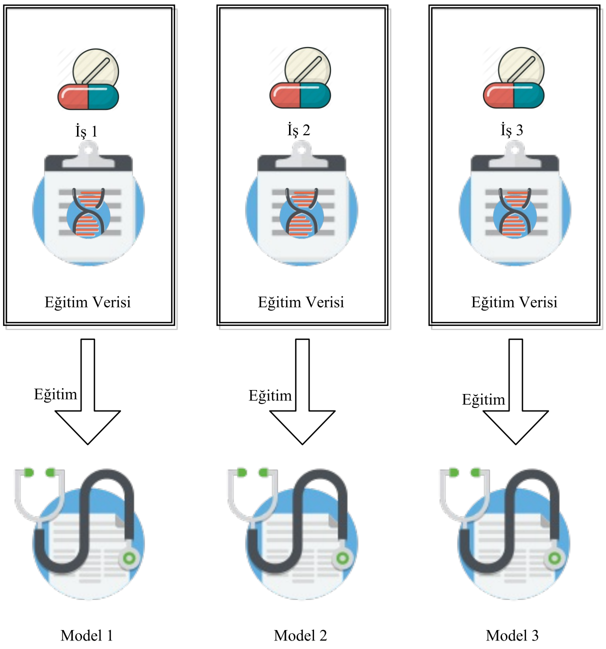
Şekil 3.2: Karar ağaçlarının birlikte kullanımı. Her bir ağacın tahmini alınarak bu değerlerin ortalaması çıktı olarak verilmiştir.

Tez çalışması sırasında oluşturulan topluluk modelinin bir parçası olan gradyan destekli regresyon için MATLAB İstatistik ve Makine Öğrenme Araç Kutusu'nda bulunan LSBoost algoritması kullanılmaktadır. LSBoost algoritmasında *Learners* parametresi ile zayıf öğrenicilerin seçimi için esneklik sağlanır. Bu parametre ile zayıf öğrenicilerin türü (ağaç, en yakın komşu veya diskriminant) belirlenir. Bu aşamada *Learners* parametresi varsayılan değişkenler ile Ağaç (Tree) olarak ayarlanmıştır.

GBR için diğer iki önemli parametre, sırasıyla modeldeki öğrenicilerin sayısı ve büzülme için öğrenme hızını belirten *NLearn* ve *LearnRate*'dır. Gradyan desteği sırasında güncelleştirme için büzülme ile regülarizasyon aşağıdaki gibidir:

$$F_t(x) = F_{t-1}(x) + \nu \cdot \gamma_t h_t(x), \quad 0 < \nu \leq 1 \tag{3.1}$$

Burada ν parametresi, öğrenme oranı olarak adlandırılır. Küçük bir öğrenme oranı



Şekil 3.3: Tek-iş öğrenme modelinin gösterimi. Tek-iş öğrenmede, her iş bağımsız olarak değerlendirilir ve bağımsız olarak öğrenilir.

seçmek modelin oluşumu için gelişim sağlar ancak hesaplama zamanı açısından da önemli bir artış olur. *NLearn* ve *LearnRate* arasındaki bu ödünleşimden dolayı bu parametreler birbiriyle ilintili olarak seçilmelidir.

3.1.2 İz-norm regülarizasyonlu çoklu-iş öğrenme

Çoklu-iş öğrenme (MTL) (Şekil 3.3), yapay öğrenme modellerinin ayrı ayrı her bir iş için eğitilmesi yerine, birbiriyle ilgili işlerin aynı anda değerlendirilmesidir. Bu şekilde farklı işler için aynı anda eğitmenin daha iyi öğrenmeye yardımcı olduğu [5] tarafından sunulmuştur. Çoklu-iş öğrenme yöntemlerinden faydalanabilmek için işler birbirleriyle ilişki içinde olmalıdır. Antikanser ilaçların modellenmesi de benzer etkiye sahip ilaçların benzer modelleri olabileceği düşüncesiyle çoklu-iş öğrenmeye uygun görülmektedir.

$$\alpha||w||_1 = \alpha \sum_i |w_i| \tag{3.2}$$

$$\alpha||w||_2^2 = \alpha \sum_i w_i^2 \tag{3.3}$$

Yapay öğrenmede oluşturulan model, eğitim verisine aşırı şekilde uyum sağladığında ve dolayısıyla modelin karmaşıklığının fazla olduğu durumlarda sıkça aşırı öğrenme problemi ile karşılaşılır. Bu problemi aşmak için ise regülarizasyon yönteminin kullanılması uygun bir çözümdür. Regülarizasyon, modelin esnekliğini azaltarak veriye aşırı uyum sağlamayı engelleyen yöntemlerin genel adıdır. Çoklu-iş öğrenme ile kullanılabilen ℓ_1-norm (3.2), ℓ_2-norm (3.3) gibi pek çok regülarizasyon algoritması vardır. Regülarizasyon kullanan makine öğrenmesi algoritmalarından biri de iz-norm regülarizasyonlu çoklu-iş öğrenmedir. Diğer regülarizasyon fonksiyonları gibi iz-norm regülarizasyon da, öğrenme modellerini, karmaşıklığı cezalandırarak gereğinden fazla uyumu önlemek için ayarlar.

$$\|W\|_p = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i^p \right)^{1/p} \tag{3.4}$$

İz-norm, aynı zamanda nükleer norm olarak da bilinir ve Schatten p-normunun (3.4) $p = 1$ olduğu yaygın bir örneğidir. Böylece iz-norm şu şekilde tanımlanabilir:

$$\|W\|_* = \text{iz}\left(\sqrt{W^*W}\right) = \sum_{i=1}^{\min\{m,n\}} \sigma_i \tag{3.5}$$

Burada σ , matrisin köşegeni üzerindeki her bir elaman olarak düşünülebilir. İz-normu temel alan TRMTL [21] ise, λ regülarizasyon parametresi olmak üzere, genel olarak şu eniyileme problemini ele alır:

$$\min_W \sum_{i=1}^{n} f(W) + \lambda \|W\|_* \tag{3.6}$$

Burada $f(W)$ yani maliyet fonksiyonunun hesaplanmasında ise en küçük kareler yöntemi kullanılır ve çoklu-iş öğrenme yöntemi için verilen (3.7) problem çözülür.

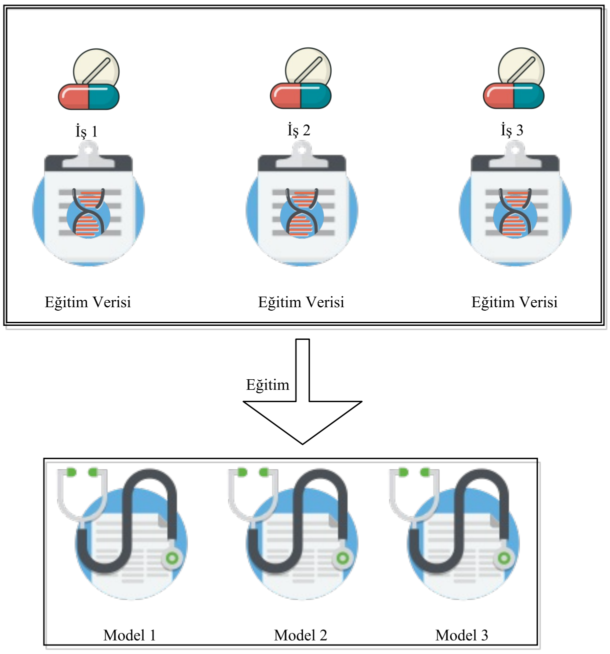
$$\min_W \sum_{i=1}^{n} \|W_i^T X_i - Y_i\|_F^2 + \lambda \|W\|_* \tag{3.7}$$

MALSAR, Multi-tAsk Learning via StructurAl Regularization, pek çok regülarizasyon algoritmasıyla birlikte çoklu-iş öğrenme yöntemlerinin uygulamasını sağlayan bir araçtır. Tez kapsamında iz-norm regülarizasyonlu çoklu-iş öğrenme metodunu da içeren bu araç kullanılarak modeller oluşturulmuştur.

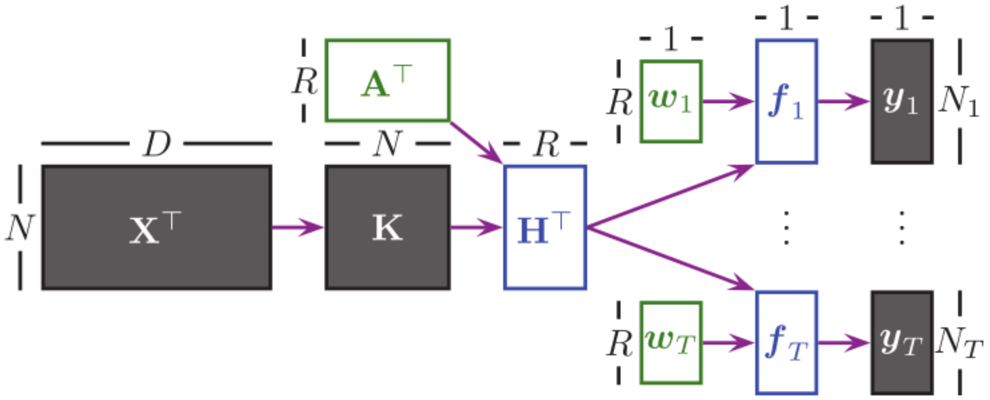
3.1.3 Çekirdekli bayes çoklu-iş öğrenme

Çekirdek(kernel) metotları, bir benzerlik fonksiyonu kullanarak veriler arasındaki ilişkileri açığa çıkaran örüntü analizi algoritmasıdır. Yapay öğrenmede çoğu zaman orijinal veri üzerinden sınıflandırma ya da regresyon algoritmalarını kullanmak yerine çekirdek metodu kullanılarak önişlemeden geçirilmiş veri kullanılır. Bu hem verinin boyutunu düşüreceği için daha hızlı çalışmasını sağlar hem de doğrusal olmayan çekirdek kullanıldığında, veriler arasındaki doğrusal olmayan ilişkileri açığa çıkarır. **X** girdi uzayındaki her bir **x** ve **x**′ için çekirdek benzerlik fonksiyonu genel olarak şu şekilde verilebilir :

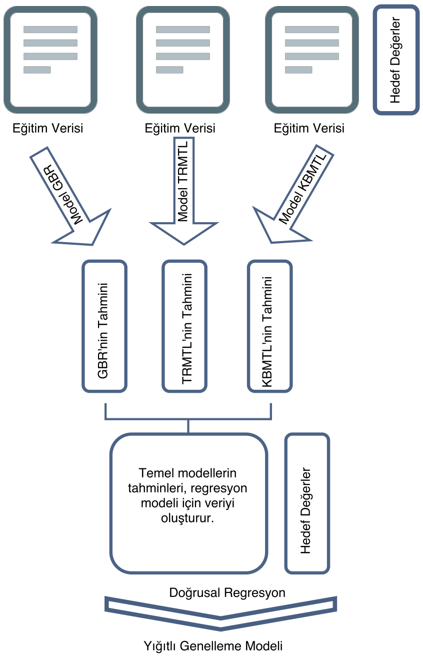
$$K(\mathbf{x}, \mathbf{x}') = \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle_{\mathcal{V}} \tag{3.8}$$



Şekil 3.4: Çoklu-iş öğrenme modelinin gösterimi. Çoklu-iş öğrenmede, işler arasındaki ilişki değerlendirilerek aynı anda birden fazla iş öğrenilir.



Şekil 3.5: İkili sınıflandırma için çekirdekli bayes çoklu-iş öğrenme (KBMTL) akış şeması.[18]



Şekil 3.6: Topluluk modelinin oluşturulması aşamasında kullanılan yığıtlı genelleme yönteminin ilk adımının şematik gösterimi.

Burada $\varphi(\mathbf{x})$, yani $\mathbf{x}$ in temsilinin nasıl olduğunun açıkça belirtilmesine gerek yoktur ancak, $\mathbf{x}$ ve $\mathbf{x}'$ iççarpım için uygun olmalıdır. Ayrıca doğrusal olmayan benzerlik fonksiyonlarıyla birlikte kullanılarak veriler arasındaki doğrusal olmayan ilişkiler de yakalanarak daha verimli algoritmalar tasarlanır. Radyal temelli fonksiyon(3.9), Gönen'in çalışmasında[18] da kullanılan doğrusal olmayan bir benzerlik fonksiyonudur.

$$K(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||^2}{2\sigma^2}\right) \tag{3.9}$$

Radyal temelli fonksiyonun σ parametresinin seçimi önemlidir. Bu parametre veri noktaları arasındaki ikili Öklid uzaklığının ortalaması kullanılarak hesaplanan bir değer ve bu değere komşu diğer dört değer üzerinden yapılan bir iç geçerleme ile belirlenmiştir.

[18] tarafından verilerde bulunan gürültüyü gidermek ve eksik değerli verileri kullanamama problemini çözmek için her yapay öğrenme işi için ortak bir altuzay kullanımı gibi yenilikleri içeren bir metot önerilmiştir. Çekirdekli bayes çoklu-iş öğrenme (KBMTL) adı verilen bu yöntem hem ikili sınıflandırma (Şekil 3.5) hem de regresyon problemleri için uygundur.

(Şekil 3.5)'de gösterildiği gibi öncelikle X verisi kullanılarak K çekirdek matrisi hesaplanır. Daha sonra A izdüşüm matrisi kullanılarak çekirdek matris bir altuzaya izdüşülür ve H gizli temsil matrisi bulunur. İkili sınıflandırma kısmında ise gizli temsil üzerinden tahmin yapılır ve bu tahminler sınıf etiketleriyle eşleştirilir.

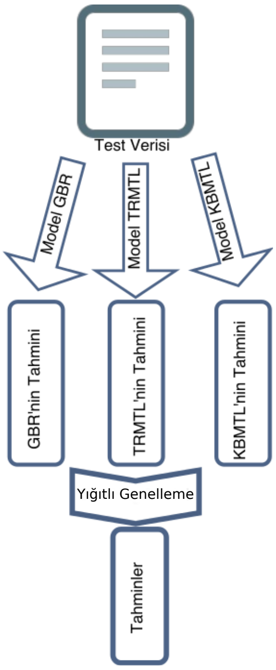
3.2 Topluluk Modeli

Toplu öğrenme, daha iyi tahminler elde edebilmek için çeşitli makine öğrenme algoritmalarının birleştirilerek yeni bir model oluşturma yöntemidir. Ortalama alma, oylama, yığıtlama gibi çeşitli topluluk modelleri vardır [31]. Tez çalışmasında yığıtlı genelleme, ortalama alma yöntemi ile karşılaştırılmış ve yığıtlı genellemenin daha iyi sonuçlar verdiği gözlenmiştir. Dolayısıyla bu çalışmada, Bölüm 3.1'de açıklanan temel modeller kullanılarak farklı birleştirme yöntemleri arasından yığıtlı genelleme (stacked generalization) [37] seçilerek topluluk modeli oluşturulmuştur.

Yığıtlı genelleme yönteminde, tahminlerin doğrusal kombinasyonlarının üretilmesi için temel tahmin modellerinin çıktıları ve eğitim verileri için hedef değerler kullanılır. Temel modellerin tahminleri öznitelik vektörü olarak kullanılarak belirlenen katsayıları içeren yığıtlama modeli, daha sonra bu modellerin sonuçlarını birleştirme işleminde kullanılır. Burada katsayısı daha yüksek olan modelin topluluk modelinin genel tahminine etkisinin daha yüksek olması beklenir.

Bu çalışmada yığıtlı genelleme oluşturulurken hem doğrusal regresyon hem de regresyon ağacı denenmiştir ve daha iyi sonuç alınan doğrusal regresyon kullanılmıştır. Ayrıca temel modellerden daha iyi performans almak için eğitim verisi (genelde kullanılan) 2-katlı yerine 5-katlı olarak ele alınmıştır. Burada verinin 5-katlı olarak ele alınması, 5 kere tekrar eden ve her seferinde verinin %80'lik bir kısmı ile modelin eğitilmesini ve kalan %20'lik kısım ile test edilip sonuçların alınmasını ifade eder.

Yığıtlı genelleme, yapay öğrenme modelinde bulunan her iş için ayrı ayrı oluşturulmaktadır. Yani aşağıda verilen adımlar takip edilerek eğitim verisi üzerinden her bir iş için ayrı yığıtlı genelleme modeli oluşturulur. Daha sonra test kısmında her örnek için, temel modellerden alınan tahminler bu yığıtlı genelleme modelinin belirttiği katsayılar kullanılarak hesaplanır.



Şekil 3.7: Yığıtlı Genelleme yönteminin model tahminlerini birleştirme adımının şematik gösterimi.

Temel alınan modelleri uygun şekilde birleştirmek için 5-katlı yığıtlı genelleme, doğrusal regresyon ile kullanıldığında adım adım aşağıdaki gibi ilerler:

- Verinin ilk %80'lik kısmıyla temel modeller oluşturularak %20'lik kısım için hedef değerleri tahmin edilir. Bu işlem 5 kere her bir kat için çalıştırılır.
- Modellerin tahminleri doğrusal regresyon modelinin girdileri, eğitim verisinin hedef değerleri ise bu oluşturulacak regresyon analizi için hedef değerler olarak kullanılır.
- Regresyon analizi sonrası öznitelik olarak ele alınan temel modellerin katsayıları belirlenir. (Şekil 3.6)
- Eğitim verisinin tamamı kullanılarak temel modeller oluşturulur.
- Temel modellerin tahminleri alınır ve belirlenen katsayılar doğrultusunda son tahmin değeri hesaplanır (Şekil 3.7).

İlaç aktivite tahmini için, topluluk yöntemi bu şekilde bir yığıtlı genelleme modeli ile kullanılarak daha güçlü bir tahmin edici oluşturulması amaçlanmıştır. Birbirlerinden farklı üç temel yöntemin yer aldığı bu topluluk modelinde, farklı ilaçlar için yöntemlerin etki katsayıları kendi içlerinde belirlenmiştir.

4. İLAÇ İMZALARINI KULLANARAK AKTİVİTE TAHMİNİ

Bu bölümde, öncelikle oluşturulan modelin karşılaştırıldığı Lasso çoklu-iş öğrenme modeline yer verilmiştir. Daha sonra imza benzerlik tabanlı modeli elde etmek için kullanılan ilaç aktivite imzalarının nasıl oluşturulduğu ve bu imzalar kullanılarak hesaplanan benzerlik ilişkisi ele alınmıştır. Son olarak ise benzerlik ilişkisinin seyrek yapı regülarizasyonlu öğrenme modeline eklenerek oluşturulan imza benzerlik tabanlı regülarizasyonlu çoklu-iş öğrenme modeli anlatılmıştır.

4.1 Lasso Çoklu-İş Öğrenme

Lasso çoklu-iş öğrenme modeli imza benzerlik tabanlı modelin karşılaştırılması için kullanılan, modelin performans ölçümü için referans alınan yöntemdir. Karşılaştırılan iki model arasındaki tek fark benzerlik ilişkisinin kullanılıp kullanılmaması olmuştur. İmza benzerlik tabanlı regülarizasyonlu çoklu-iş öğrenme modeli, MALSAR tarafından sağlanan Lasso çoklu-iş öğrenme 4.1 modeli ile karşılaştırılarak benzerlik ilişkisinin etkisi ölçülmüştür.

$$\min_{W} \sum_{i=1}^{n} \|W_i^T X_i - Y_i\|_F^2 + \rho_2 \|W\|_1 \tag{4.1}$$

Bu modelde adından da anlaşılacağı üzere aşırı öğrenmeyi önleme ve öznitelik seçimi için Lasso regülarizasyonu kullanılmıştır. ℓ_1 norm olarak da bilinen bu regülarizasyon yöntemi model katsayılarını seyrekleştirmeyi amaçlamaktadır. Bu model ayrıca çoklu-iş öğrenme yöntemi ile Lasso regülarizasyonunu bir arada kullanarak, işler için tüm model katsayılarının eş zamanlı olarak belirlenmesini sağlar.

4.2 İlaç Aktivite İmzasının Oluşturulması

LINCS [13] veri tabanında bulunan ilaç-hücre hattı deneylerine, lincscloud uygulama programlama arayüzü(API) [1] aracılığıyla erişim sağlanabilmektedir. Bu deneyler, ilaçların hücre hatlarındaki hangi genleri nasıl değiştirdiği bilgisini içerir. Lincscloud API' nin, verilen ilaç-hücre hattı çifti için regüle edilmiş probe kümesini sağladığı göz önünde bulundurularak, bir ilacın aktivite imzası oluşturulurken o ilacın veri tabanında bulunan tüm hücre hattı deneyleri sorgulanmıştır.

---

[1] http://api.lincscloud.org/

İlaç etkisiyle değişen probe kümelerine, veri tabanından farklı şekillerde erişilebilir. Tez çalışmasında ilaç aktivite imzası oluşturulurken, bu adımda belirli 1000 probe'tan en çok değişen 50 tanesini içeren küme hesaba katılmıştır. Veri tabanı küratörleri tarafından belirlenen 1000 probe'luk bu liste, gen ifadesindeki bilgiyi genel olarak özetlemektedir[2]. Daha sonra veritabanında probe değişimi cinsinden yer alan bu bilgiler mygene [38] kütüphanesi kullanılarak gen değişimlerine dönüştürülmüştür. Bu noktada *dr* ilaç, *CL* hücre hattı olmak üzere bir (dr,CL) çifti için aşağıda verilen şekilde deney imzası (ExpSig) oluşturulmuştur:

$$ExpSig(dr,CL) = \langle gen_1 \uparrow, gen_2 \downarrow, gen_3 \downarrow, ..., gen_K \uparrow \rangle \tag{4.2}$$

Aşağıda örnek olarak doramapimod ilacının MCF7 hücre hattına uygulanması sonucunda; ACAT2 ve ADBR2 genlerinin yukarı yönlü regüle edilmiş olduğu, ADH5 ve ZMIZ1 genlerinin aşağı yönlü regüle edilmiş olduğu gösterilmiştir.

$$ExpSig(doramapimod,MCF7) = \langle ACAT2 \uparrow, ADH5 \downarrow, ADRB2 \uparrow, ..., ZMIZ1 \downarrow \rangle \tag{4.3}$$

Bir ilaç-hücre hattı çifti için deney parametrelerine bağlı olarak (ilacın ne kadar süre uygulandığı, ilaç dozu vb.) birden fazla deney bulunabilmektedir. Bu farklılaşma ile deney sayısı her çift için farklı sayıda olmaktadır. Örneğin; (vorinostat,MCF7) çifti için veri tabanında 188 deney bulunmaktadır. Bu farklılaşmış deneyler tez çalışmasında ayrı deneyler olarak ele alınmıştır ve bu deney imzaları kullanılarak verilen *dr* ilacı için, ilaç aktivite imzası (ActSig) oluşturulmuştur.

$$ActSig(dr) = \forall CL \cup_{dr} ExpSig(dr,CL) \tag{4.4}$$

Burada $\cup_{dr}$, *dr* ilacı için deney imzalarının birleşimini ifade eder. Bu noktada ilacın gene olan etkisinin keşfedilmesi için kaç kere yukarı, kaç kere aşağı yönlü değişim gösterdiği verisi kullanılmıştır.

$$ActSig(dr) = \langle g_1(\uparrow n_1, \downarrow n_1), ..., g_K(\uparrow n_K, \downarrow n_K) \rangle \tag{4.5}$$

$g_1(\uparrow n_1)$ ifadesi, *dr* ilacının g_1 genini $(\uparrow n_1)$ defa yukarı yönlü, $(\downarrow n_1)$ defa aşağı yönlü değiştirdiğini ifade etmektedir. Doramapimod ilacı için örnek bir imza aşağıdaki gibi gösterilebilir:

$$ActSig(doramapimod) = \langle MRPS2(1,19), MEF2C(12,4), ..., SRC(20,2) \rangle \tag{4.6}$$

Burada, tüm deneylerde zebularine ilacı için, MRPS2 geninin 1 defa aşağı yönlü, 19 defa yukarı yönlü değişime uğradığı ifade edilmektedir.

---

[2] http://support.lincscloud.org/hc/en-us

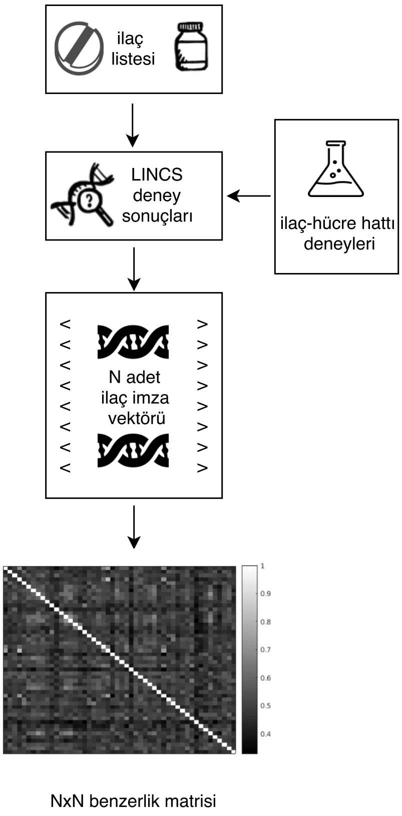
4.3 İlaç Etki Benzerliklerinin Hesaplanması

Benzerlik hesaplanması için ilaçların genler üzerindeki etkisinin keşfedildiği aktivite imzaları kullanılırken, benzer ilaçların benzer genleri etkilediği hipotezi esas alınmıştır. Bir ilacın aşağı ya da yukarı yönlü en çok değiştirdiği belirli bir sayıdaki gen listesi ile bir başka ilacın aynı şekilde değiştirdiği gen listesi karşılaştırılarak, bu listelerin kesişimi ne kadar çok ise ilaçlar o kadar benzerdir yorumu yapılmıştır. Bu bilgiler ışığında iki ilacın benzerliği aşağıdaki gibi hesaplanmaktadır:

$$liste_{dr} = maxN(ActSig(dr)) \tag{4.7}$$

$$Benzerlik(dr_A, dr_B) = \frac{|liste_{dr_A} \cap liste_{dr_B}|}{N} \tag{4.8}$$

Burada önemli nokta kesişimleri alınacak gen listeleri için, listelerin uzunluğunun, N, belirlenmesidir. Bu uzunluk, alınan farklı uzunluklardaki listeler ile yapılan çalışmaların sonuçlarına bakılarak belirlenmiştir. Bu hesaplama tüm ikili ilaç kombinasyonları için yapılarak benzerlik matrisi Şekil 4.1'teki gibi oluşturulmuştur.



Şekil 4.1: İlaç aktivite imzalarının kullanılarak benzerlik matrisinin oluşturulması

Şekil 4.1'de gösterildiği gibi veri kümelerinden sağlanan ilaç listeleri için imza üretmek üzere, ilaçların LINCS üzerinde bulunan deneyleri sorgulanır. Her bir deneyin sonuçlarına bakılarak, o ilacın etkisiyle aşağı ya da yukarı yönde regüle edilmiş genler belirlenir. Belirlenen genler ve regüle edilme sayıları ilacın imzasını oluşturur. Daha sonra ilaçların benzerlikleri içerdeki ortak gen sayısına bakılarak belirlenir.

4.4 İmza Benzerliği Tabanlı Regülarizasyonlu Çoklu-İş Öğrenme

Çoklu-iş öğrenme metotlarında öğrenilen işlerin birbirleriyle ilişkili olması gerekir. Ancak çoğu durumda verilen işlerin tamamının birbiriyle ilişkili olması beklenemez. Örnek olarak; bazı kanser ilaçları birbirleriyle ilişkili olabilirken, bazı ilaç çiftleri arasında bir ilişki bulunmuyor olabilir. Bu gibi durumlarda da çoklu-iş öğrenme metotlarından iyi bir şekilde yararlanılması için, benzerlik ilişkileri girdi olarak alınarak, öğrenilen modellerin katsayılarının regülarizasyonunda bu benzerlikten yararlanılır. Bu regülarizasyon yönteminde bağlantılı olduğu düşünülen iş çiftlerinin model katsayıları farkı azaltılmaya çalışılır.

Bu çalışmada kullanılmak üzere MALSAR [44] kütüphanesinde bulunan seyrek çizge regülarizasyonlu çoklu-iş öğrenme modeli seçilmiştir. Bu modelin girdi olarak aldığı çizge yapısı benzerlik ilişkilerini taşıyacak şekilde modele verilmektedir. MALSAR'da bulunan SRMTL metodu; aşağıda verilen çizge yapı, ℓ-1 norm ve ℓ-2 norm regülarizasyonlu problemi ele alır:

$$\min_{W} \sum_{i=1}^{n} \|W_i^T X_i - Y_i\|_F^2 + \rho_1 \|WR\|_F^2 + \rho_2 \|W\|_1 + \rho_{L2} \|W\|_F^2 \tag{4.9}$$

Burada W_i, X_i, Y_i sırasıyla; i numaralı işin modelini, i numaralı girdiyi ve i numaralı işin hedef değerlerini belirtir. ρ_1, ρ_2 ve ρ_{L2} ise model katsayılarının seyrekliğini kontrol eden regülarizasyon parametreleridir. ρ_2 ve ρ_{L2} parametreleri isteğe bağlı olarak modele verilebilir. Çalışmalarımızda, bu iki parametreden ℓ-1 norm için gerekli olan ρ_2 parametresi, çizge seyrekliğini kontrol eden ρ_1 parametresi ile birlikte kullanılarak model oluşturulmuştur.

Problem 4.9'de verilen R parametresi ise iş benzerlik ilişkilerinin çizge (graph) üzerinde temsil edilmesini sağlar. Bu gösterimde; işlerin her biri birer düğüm (node) olarak düşünülür ve eğer iki iş arasında bir benzerlik varsa, bu iki iş birbirine bir kenar (edge) ile bağlıdır. k kenar olmak üzere, i numaralı kenar için, A ve B işleri birbirine bağlı ise bu kenar;

$$k_A^{(i)} = \sqrt{Benzerlik(dr_A, dr_B)} \tag{4.10}$$

ve

$$k_B^{(i)} = (-1) * \sqrt{Benzerlik(dr_A, dr_B)} \tag{4.11}$$

şeklinde bir vektör ile gösterilmiştir. K bütün kenar kümesini simgelerse R çizgesi şu vektörlerden oluşmuştur:

$$R = [k^{(1)}, k^{(2)}, k^{(3)}, \ldots, k^{\|K\|}] \in \mathbb{R}^{tx\|K\|} \tag{4.12}$$

Bu bilgiler ile birlikte, kenarlardan oluşan R'nin kullanıldığı $\|WR\|_F^2$ ifadesini daha açık bir biçimde yazmak gerekirse :

$$\|WR\|_F^2 = \sum_{i=1}^{\|K\|} \|Wk^{(i)}\|_2^2 = \sum_{i=1}^{\|K\|} \|W_{k_A^{(i)}} - W_{k_B^{(i)}}\|_2^2 \tag{4.13}$$

Dolayısıyla bu regülarizasyon işlemi ile benzer işlerin model kat sayıları arasındaki farkın azaltılması sağlanır. Eğer işler birbirine benzemiyorsa kat sayılar arasındaki fark önemsizdir. Bu durum, R üzerindeki ilgili indeks değerleri sıfır yapılarak sağlanmıştır. İşlerin benzer olduğu durumda ise R üzerindeki ilgili indeks değerleri yüksek değerler olacağı için, algoritma tarafından model katsayıları arasındaki fark küçültülmeye çalışılır.

Regülarizasyon parametresi olan ρ_1'in belirlenmesi aşamasında, parametreyi seçmek için grid arama algoritması kullanılmıştır.

5. DENEYSEL SONUÇLAR

Öncelikle yapılan deneylerde kullanılan modeller için belirlenen parametreler ve bu parametrelerin nasıl belirlendiği Bölüm 5.1'de ele alınmıştır.

Bölüm 5.2'de anlatılan GDSC ve CTRP veri kümelerinden sağlanan ilaç-hücre hattı aktivite değerleri ve hücre hattı gen ifade profilleri kullanılarak temel alınan modeller ile bu modellerin birleşimiyle oluşturulan topluluk modelinin tahmin güçleri ölçülmüştür. Bu ölçüm öncesi veriyi hazırlamak üzere bazı önişleme teknikleri kullanılmıştır. Bu teknikler 5.3 bölümünde anlatılmıştır. Daha sonra hazırlanan veri ile oluşturulan modellerin tahmin güçlerinin karşılaştırılması için çapraz doğrulama kullanılmıştır.

Tahmin güçlerine bakıldığında oluşturulan topluluk modelinin ilaç-hücre hattı aktivitesini diğer modellerden daha iyi tahmin ettiği sonucuna varılmıştır. Sonucun ışığında 5.5 bölümde ise veri kümelerinde eksik olarak belirtilen ilaç-hücre hattı çiftleri için tahminler yapılmıştır.

Bölüm 5.6'da ise imza benzerlik tabanlı çoklu-iş öğrenme modeli için çapraz doğrulama sonuçları yer almaktadır. Burada oluşturulan model yine GDSC ve CTRP veri kümeleri kullanılarak test edilmiştir. Oluşturulan modelin performansının karşılaştırılması için referans alınan modelin sonuçları da bu kısımda yer almaktadır.

5.1 Ayarlar

Topluluk yönteminin oluşturulması kısmında tamamen MATLAB yazılımı kullanılmıştır. İlaç benzerlik tabanlı yöntemde ise imza oluşturma ve benzerlik hesaplama aşamasında Python programlama dilinden yararlanılmış, modellerin oluşturulması ve değerlendirilmesinde ise yine MATLAB kullanılmıştır.

Topluluk modeli oluştururken kullanılan temel modeller olan, gradyan destekli regresyon, iz-norm regülarizasyonlu çoklu-iş öğrenme ve çekirdekli bayes çoklu-iş öğrenme yöntemleri için model parametreleri ayrı ayrı belirlenmiştir.

Tez kapsamında GBR için, modeldeki öğrenici sayısının belirlendiği *NLearn* ve büzülme için öğrenme hızını belirten *LearnRate* için farklı seçenekler denenmiş ve bu parametreler 100 ve 0.1 olarak ayarlanmıştır . Tanımlanan parametrelerle GBR kullanılarak, iyi bir tahmin modeli elde edilmiştir.

İz-norm için ise regülarizasyon parametresi olan λ'yı eniyilerken eğitim verisi üzerinde 5 katlı grid arama algoritması kullanılmıştır. [0.1, 1, 10, 100] olarak verilen liste içerisinden parametre seçimi yapılmıştır.

KBMTL ise, çok sayıda parametre isteyen bir yöntem olmasına rağmen, MATLAB uygulamasında verilen, varsayılan parametrelerle kullanılmıştır. Örneğin; alt uzay boyutsallığı 20 olarak ayarlanmıştır.

Veri kümelerinde yer alan ilaçlar için aktivite imzası oluştururken ilaçlara göre deney sayısı farklılık göstermiştir. Sonuç olarak toplam deney sayısının GDSC'de bulunan ilaçlar için 6 - 1057 aralığında, CTRP'de bulunan ilaçlar için 2 - 1000 aralığında değiştiği gözlenmiştir.

Aktivite imzası oluşturma işlemi, GDSC ve CTRP veri kümelerinde bulunan ilaçlar için ayrı ayrı uygulanmıştır. GDSC veri kümesi için 265 ilaçtan 133, CTRP için 481 ilaçtan 286 tanesi için aktivite imzası oluşturulmuştur. Tüm ilaçlar için imza oluşturulamamasının sebebi, LINCS veritabanında bulunan ilaçlar ile diğer veritabanlarındaki ilaçların eşleştirilememesi ya da LINCS'te ilaçların bulunmamasıdır. Eşleşme problemi, ilaçların farklı veri tabanları için ortak bir kimliğinin bulunmamasından kaynaklanmaktadır. İlaç aktivite imzaları, eşleşen ilaçlar için, benzerlik hesaplanmasında kullanılmak üzere hazır hale getirilmiştir.

Benzerlik hesaplamasında ise önemli nokta, kesişimleri alınacak gen listeleri için listelerin uzunluğunun, N, belirlenmesidir. Bu uzunluk, alınan farklı uzunluklardaki listeler (10, 20, 40, 80) ile yapılan çalışmaların sonuçlarına bakılmıştır ve 80 olarak belirlenmiştir.

İmza benzerlik tabanlı regülarizasyonlu çoklu-iş öğrenme modelinde kullanılan regülarizasyonların her ikisi(ℓ-1 norm ve benzerlik tabanlı) için de eğitim verisi üzerinde yapılan 5 katlı grid arama algoritması kullanılarak parametreler belirlenmiştir. Ayrıca bu yöntemin karşılaştırılmasında kullanılan Lasso çoklu-iş öğrenme modeli için de regülarizasyon parametresi belirlenirken yine grid arama algoritması kullanılmıştır.

5.2 Veri Kümeleri

Oluşturulan topluluk modelinin ve imza benzerlik tabanlı modelin tahmin güçlerinin değerlendirilmesi için hücre hattı-ilaç tepkileri ve ilgili hücre hatlarının gen ifadeleri verilerinin sağlandığı iki önemli veri kümesi (GDSC ve CTRP) kullanılmıştır. Kullanılan veri kümelerinin bazı özellikleri karşılaştırmalı olarak Şekil 5.1'de verilmiştir. Son olarak verilen LINCS veri kümesinden ise ilaç imzalarının oluşturulması aşamasında yararlanılmıştır.

Çizelge 5.1: Kullanılan veri kümelerinin bazı özellikleri

Veri Kümesi	**GDSC**	**CTRP**
İlaç Sayısı	265	481
Hücre hattı sayısı	1074	860
Tepki Sayısı	224,510	314,464
AUC	✓	✓
IC_{50}	✓	✗

5.2.1 Kanserde İlaç Hassasiyet Genomiği

Topluluk modelinin ve ilaç benzerlik tabanlı modelin değerlendirilmesi için kullanılan ilk veri kümesi kanserde ilaç hassasiyet genomiği (genomics of drug sensitivity in cancer) (GDSC) [39] kümesidir. GDSC 2012 yılında oluşturulmuştur ve düzenli olarak güncellenmektedir. Bir çok araştırmacı çalışmalarını değerlendirmek için GDSC'de bulunan verileri kullanmaktadır. Tez çalışmasında Temmuz 2016'da yayınlanan güncel sürümü, GDSC v17, kullanılmıştır. Güncellenen sürümünde GDSC, 265 ilaç, 1074 hücre hattı ve 224.510 ilaç tepki değerinden oluşur. İlaçların hücre hatları üzerindeki tepki değerleri hem doz-tepki eğrisi altındaki alan (AUC) hem de yarı maksimum durdurucu konsantrasyon değerinin doğal logaritması ($\log(IC_{50})$) cinsinden verildiği için, topluluk modelinin değerlendirme aşamasında her iki hedef değeri de kullanılmıştır. İlaç benzerlik tabanlı model için ise sadece $\log(IC_{50})$ değeri kullanılmıştır.

Ayrıca, hücre hatlarının gen ifadeleri, RMA normalleştirilmiş bazal ifade profilleri olarak sunulmuştur. Veri kümesinde yer alan hücre hatlarından bazılarının gen ifade verisinde bazı eksikler bulunmaktadır. Gen ifadesi verileri olmayan bu hücre hatları çıkartıldıktan sonra, kalan hücre hattı sayısı 1014'tür. Her ilaç için tüm hücre hatlarının deneyleri yer almadığı için toplam 1014 hücre hattı olmasına rağmen, GDSC'de ilaçlar için deneylenen hücre hattı sayısı ise 363(Rapamycin için) ve 940(Bleomycin-50uM için) aralığında değişkenlik gösterir.

5.2.2 Kanser Tedavi Tepki Portalı

Modellerin değerlendirilmesinde kullanılan diğer veri kümesi de kanser tedavi tepki portalıdır. (cancer therapeutics response portal) (CTRP) [28, 30].CTRP'nin ilk sürümü 2012 yılında yayınlanmıştır. Tez çalışmasında kullanılan sürüm ise 2015 yılında yayınlanan ve 2016 yılında bazı önişleme işlemleri ile güncellenen sürümüdür. CTRP veri kümesinin güncel sürümünde 481 ilaç ve 860 hücre hattı bulunmaktadır. Bu veri kümesinde ilaç-hücre hattı çiftlerinin duyarlılık skorları sadece doz-tepki eğrisi altındaki alan (AUC) cinsinden verilmiştir. Hücre hatları için verilen gen ifadesi değerleri için ise log2 dönüşümü yapılmış ortalama değerler kullanılmıştır. Gen ifadesi değerleri, AUC ile eşleştirilerek modellerin değerlendirilmesinde kullanılmıştır.

GDSC'dekine benzer olarak CTRP verisinde de bazı hücre hatlarının gen ifadesi verileri eksiktir. Bu hücre hatları veri kümesinden çıkartılarak, 823 hücre hattı üzerinden çalışmalar yapılmıştır. Ayrıca bazı ilaçlar için yetersiz sayıda hücre hattı tepkisi olduğu belirlenmiştir. Bu sebeple belli bir değerin altında deneye sahip ilaçların modelde kullanılmaması gerektiği çıkarımıyla yeterli örneğe sahip ilaçlar 439 adet olarak belirlenmiştir. Bu 439 ilaç arasında en az ve en çok deney sonucuna sahip ilaçlar sırasıyla, MG-132(299 deney sonucu) ve Leptomycin(809 deney sonucu)'dir.

5.2.3 Tümleşik Ağ Tabanlı Hücresel İmza Kütüphanesi

Bir hücrenin fenotipinin belirli etmenler tarafından nasıl ve ne zaman değiştirildiğini gözlemlemek, hastalığa karışan mekanizmalar hakkında ipucu sağlayabilir. Tümleşik ağ tabanlı hücresel imza kütüphanesi (LINCS) projesi de, bir biyolojik işlemin herhangi birinin bozulmasının, hücrenin moleküler ve hücresel özelliklerinde, davranışında ve işlevinde değişikliklere neden olacağı öncülüne dayanmaktadır. LINCS veri kümesi, kimyasal bileşikler ile tedavi edilen insan hücrelerinin test sonuçlarını içerir. Ayrıca LINCS verileri, bilim insanlarının güncel hastalıklar,ilaçlar ve tedavi yöntemleri ile ilgili problemleri ele almasını kolaylaştırmak için topluluk kaynağı olarak açıkça kullanılabilir hale getirilmiştir. GDSC ve CTRP veri kümelerinde yer alan ilaçlar için imza oluşturulmak amacıyla LINCS veri kümesinde bulunan test sonuçları kullanılmıştır. LINCS veri kümesinde deneylenen çok sayıda bileşik bulunurken toplamda yalnızca 76 hücre hattı ile bu ilaç deneyleri yapılmıştır.

5.3 Veri Önişleme

Veri önişleme tekniği, veri madenciliği için önemli bir konudur. Normalizasyon, standartlaştırma, gürültü temizleme, öznitelik seçimi gibi teknikler verinin bir sonraki işlem için hazırlanmasını sağlar. Tez çalışmasında aktivite verileri için herhangi bir önişleme yapılmazken, gen ifadeleri için öznitelik seçimi, standartlaştırma ve çekirdek yöntemi kullanılarak boyutsal küçültme yapılmıştır.

5.3.1 Öznitelik seçimi

Gen ifadesi profilleri hücre hatlatı için kullanılan her iki veri kümesi tarafından da ayrı ayrı sağlanmıştır. Bu gen ifadesi verileri çok boyutlu bir veridir, GDSC'den elde edilen veri kümesinde 17.737, CTRP'de ise 18.541 farklı gen için sonuçlar verilmiştir. Bu hem gen sayısı, yani öznitelik sayısı, örnek sayısını aştığı için, hem de bu genlerin tamamının kanser ile ilişkisi olmadığından genler arasında bir seçim yapma ihtiyacı doğmuştur.

Bu ihtiyaç doğrultusunda Malacards [27] veri tabanından yararlanan bir gen seçim prosedürü uygulanmıştır. Bu veritabanını kullanarak, belirli bir hastalıkla ilişkili olduğu bilinen genlerin güncel listesi alınabilir. Buradan yola çıkılarak hücre hatlarının kanser türlerini temel alan bir dizi anahtar kelime üretilmiştir ve kanser hücre hatları ile ilgili 1545 genden oluşan liste indirilmiştir. Tüm modeller için öznitelik olarak, kullanılan veri kümelerinin gen ifadesi verilerindeki gen listesi ile bu listenin kesişiminde yer alan genlerin listesi kullanılmıştır.

5.3.2 Standartlaştırma

Standartlaştırma öznitelikler arasındaki birim farkını ortadan kaldırmak için uygulanan bir tekniktir. Standartlaştırma sonucunda ortalaması sıfır olan ve birim standart sapmaya sahip veri elde edilir. Eğitim verisi için; her bir değerin eğitim verisinin ortalaması ile farkı alınır ve eğitim verisinin standart sapmasına bölünür. Test verisi için ise; test verisindeki değerler ile eğitim verisinin ortalamasının farkı alınır ve eğitim verisinin standart sapmasına bölünür. Bunun için z skor formülü kullanılmıştır:

$$z = \frac{x - \mu}{\sigma} \tag{5.1}$$

burada μ eğitim verisinin ortalaması, σ ise eğitim verisinin standart sapmasıdır.

Bu standartlaştırma işlemi gen ifadesi verileri üzerinde uygulanmıştır. Topluluk modeli, TRMTL ve KBMTL algoritmalarında standartlaştırılmış veriler kullanılmıştır. GBR gibi ağaç tabanlı algoritmalar için ise standartlaştırma işlemine gerek yoktur çünkü bu algoritma sadece verinin daha büyük ya da daha küçük olması ile ilgilenir. İmza benzerlik tabanlı çoklu-iş öğrenme yönteminde ise TRMTL ve KBMTL'de olduğu gibi standartlaştırılmış veriler kullanılmıştır.

5.3.3 Boyutsal küçültme

Boyutsal küçültme işlemi verinin öznitelik sayısının azaltılmasını sağlayan bir tekniktir. Bu işlem için temel bileşenler analizi yönteminden faydalanılabileceği gibi, çekirdek numarası kullanılarak hem boyutta azaltma sağlanır hem de doğrusal modeller, doğrusal olmayan boyuta taşınır. Bu amaçla radyal temelli fonksiyon (RBF) çekirdeğin kullanılması bilinen bir yöntemdir.

Gönen de çalışmasında [18] yine bu boyutsal küçültme işlemini önermiştir. Buradan yola çıkılarak tez çalışmasının ilk kısımındaki topluluk yöntemi için hem KBMTL hem de TRMTL'de RBF çekirdeği kullanılarak bu boyutsal küçültme işlemi uygulanmıştır. GBR için ise standartlaştırma işleminde olduğu gibi bu teknik de göz ardı edilmiştir. İkinci kısımdaki imza benzerlik tabanlı model için de yine RBF çekirdeği kullanılmıştır.

5.4 Topluluk Modeli İçin Çapraz Doğrulama

Topluluk modelinin diğer yöntemlerden daha iyi performans gösterdiğinin doğrulanması için, öğrenme modelleri 10-katlı çapraz doğrulama ile değerlendirilmiştir. Performans karşılaştırması üç farklı ölçev kullanılarak yapılmıştır. Bu ölçevler; ilaçların ortalama karesel hatalarının (5.2) ortalaması (AMSE) (5.3), ilaçların ortalama karesel hatalarının ağırlıklı ortalaması (WAMSE) (5.4) ve her tahmin edicinin en iyi olarak tahmin ettiği ilaç sayısı (NDPB)'dır. AMSE ve WAMSE aşağıdaki gibi hesaplanmıştır.

$$\text{MSE} = \frac{1}{n}\sum_{i}^{n}(\hat{Y}_i - Y_i)^2 \tag{5.2}$$

$$\text{AMSE} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n_t}\sum_{i=1}^{n}(\hat{Y}_{t_i} - Y_{t_i})^2 \tag{5.3}$$

Çizelge 5.2: Topluluk Yöntemi için GDSC veri kümesi(IC_{50}) sonuçları

	AMSE	**WAMSE**	**NDPB**
GBR	1.65	1.63	26 / 265
TRMTL	1.87	1.87	4 / 265
KBMTL	1.83	1.81	6 / 265
Topluluk	**1.60**	**1.58**	**229** / 265

Çizelge 5.3: Topluluk Yöntemi için GDSC veri kümesi(AUC) sonuçları

	AMSE	**WAMSE**	**NDPB**
GBR	1.51×10^{-2}	1.51×10^{-2}	**160**/ 265
TRMTL	1.84×10^{-2}	1.85×10^{-2}	18 / 265
KBMTL	1.72×10^{-2}	1.71×10^{-2}	4 / 265
Topluluk	1.51×10^{-2}	1.5×10^{-2}	83 / 265

$$\text{WAMSE} = \frac{1}{\sum_{t=1}^{T} n_t} \sum_{t=1}^{T} \sum_{i=1}^{n} (\hat{Y}_{t_i} - Y_{t_i})^2 \tag{5.4}$$

Burada T ilaç sayısını, n ise her bir ilacın örnek sayısını belirtir. İlaç hatalarının ortalamasının yanı sıra ağırlıklı ortalamasının verilmesinin sebebi, ilaçların eşit sayıda örnek sayısına sahip olmamasıdır. Yani buradaki ağırlıklar ilaçların deneylendiği hücre hattı sayısıdır.

İlaçların aktivite değerlerinin tahmininde hatalar hesaplanırken MSE ölçütü kullanılmıştır. Daha sonra modeller birden fazla ilacın tahmininde kullanıldığı için bu ilaç hatalarının ortalaması alınarak modellerin ilaçların geneli üzerindeki tahmin gücü, modellerin karşılaştırılmasında kullanılmıştır. Burada NDPB ölüçütünde ise bir ilacı tahmin eden en iyi modeli bulmak için, o ilaç üzerindeki MSE değeri en küçük olan model seçilmiştir.

GDSC veri kümesinde ilaçların hücre hatları üzerindeki hassasiyeti iki farklı şekilde verilmiştir; IC_{50} ve AUC. Dolayısıyla tez çalışmasında da her iki değer üzerinden deneyler yapılmıştır. GDSC verisi için öncelikle hedef olarak belirtilen IC_{50} değerleri tahmin edilmeye çalışılmıştır. Temel olarak alınan üç modelin ve bunların birleşimiyle oluşturulan topluluk modelinin bu değeri tahmin etmedeki gücü Çizelge 5.2'de gösterilmiştir. Bu çizelgede topluluk modelinin tüm ölçütlerde diğer modellerden daha iyi olduğu görülmektedir. Toplam 265 ilaçtan 229 tanesinde en iyi tahmin perfonmasına sahip model topluluk modeli olmuştur.

Çizelge 5.3'de ise yine GDSC veri kümesi kullanılarak, hedef olarak AUC değerleri tahmin edilmiştir. Bu tahmin sonuçlarında belirgin bir üstünlükten söz edilemez; topluluk modeli, GBR ile birlikte en iyi performansı göstermiştir.

CTRP veri kümesi ilaçların hücre hatları üzerindeki aktivitesini AUC değerleri üzerinden paylaşmıştır. Bu sebeple CTRP için deneylerde hedef olarak AUC değerleri kullanılmıştır. Bu değerin tahmini Çizelge 5.4'de modeller arasındaki performansların karşılaştırılması için verilmiştir. Yapılan değerlendirmede topluluk modelinin CTRP

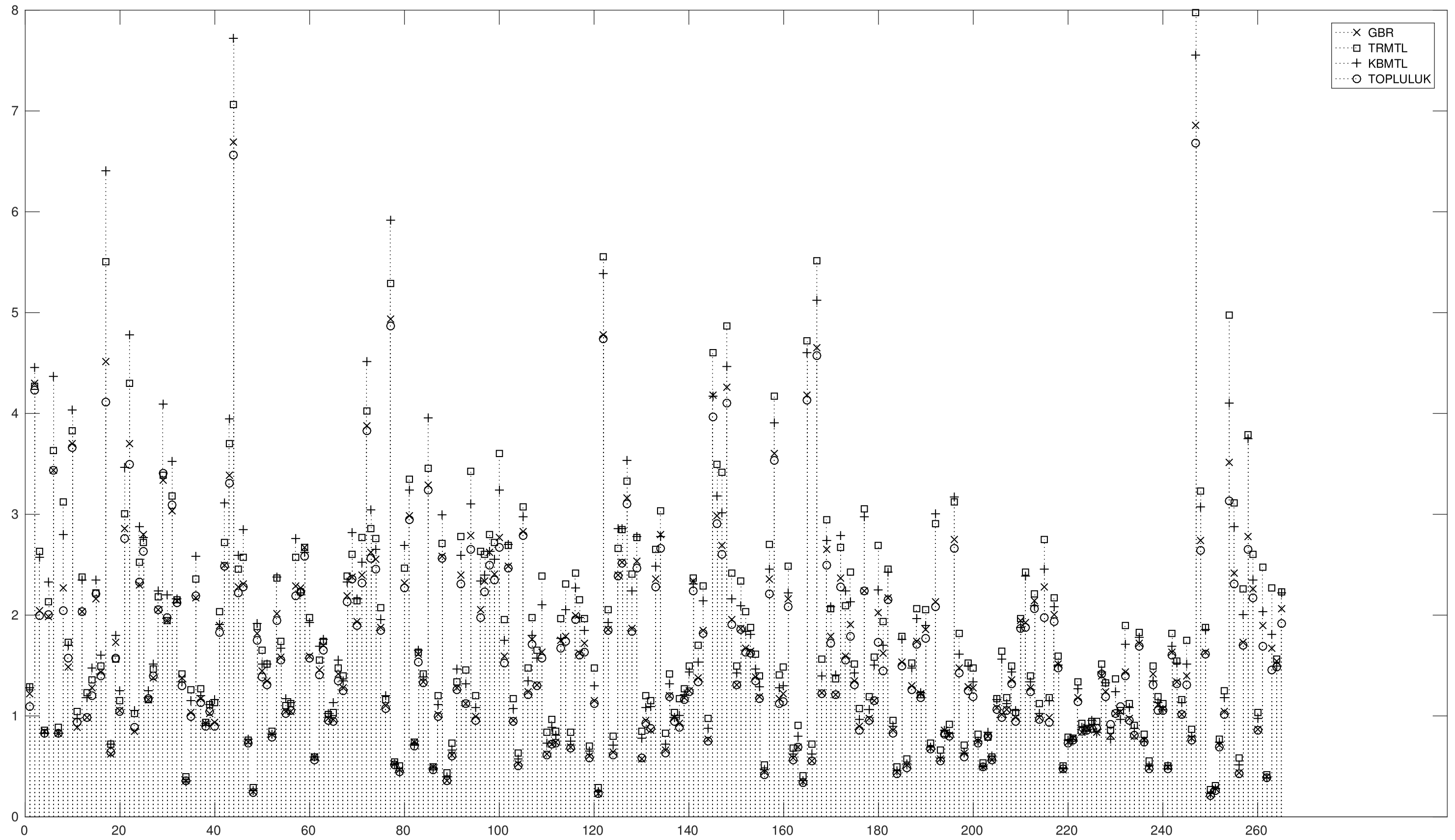
Çizelge 5.4: Topluluk Yöntemi için CTRP veri kümesi(AUC) sonuçları

	AMSE	**WAMSE**	**NDPB**
GBR	2.07	2.09	68 / 439
TRMTL	2.38	2.40	63 / 439
KBMTL	2.32	2.33	10 / 439
Topluluk	**2.03**	**2.05**	**298** / 439

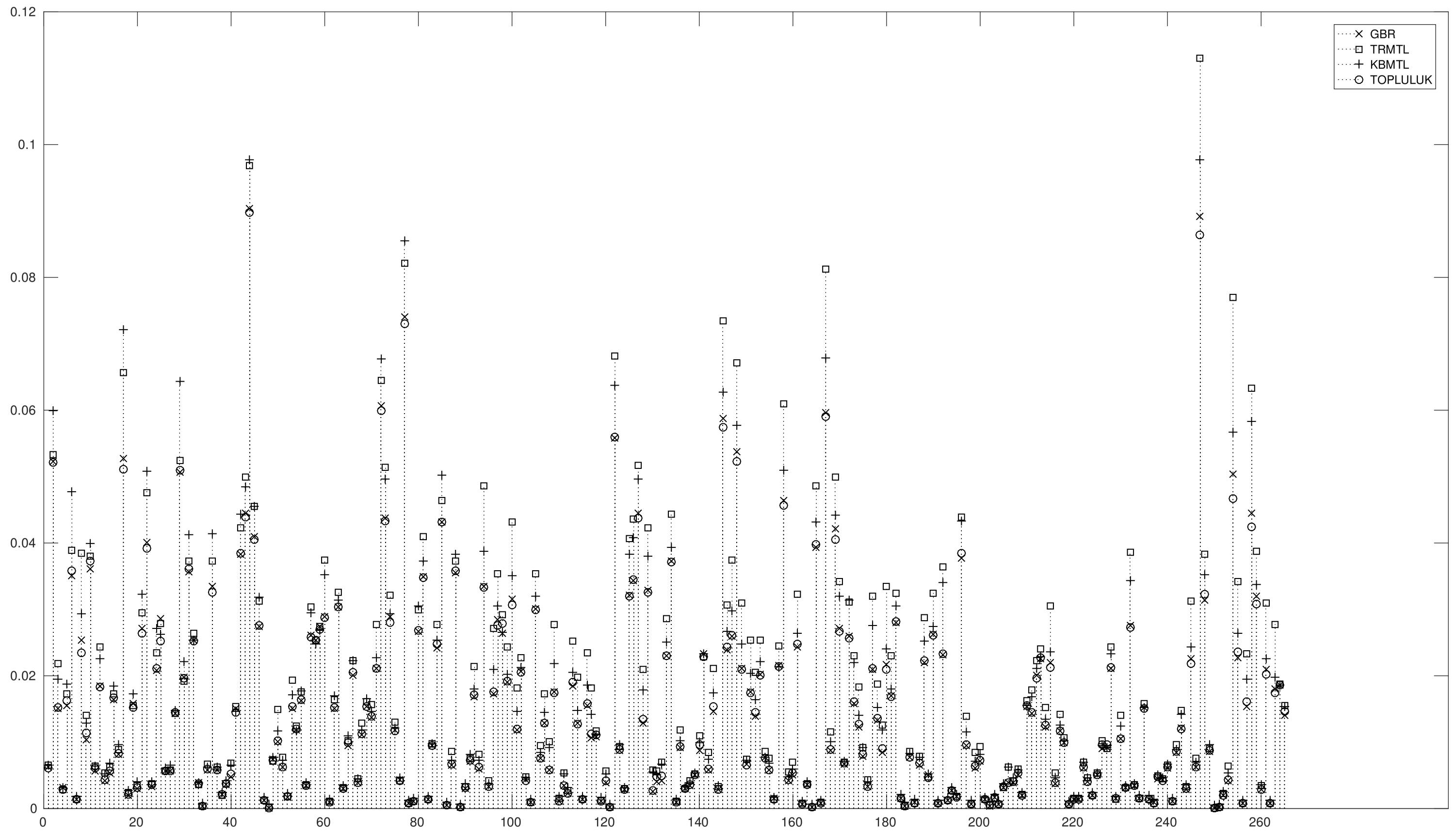
veri kümesi için de olumlu sonuçlar verdiği görülmüştür. Veri kümesinde bulunan 439 ilaçtan 298 tanesi için topluluk modelinin performansının en iyi olduğu gözlenmiştir.

Şekil 5.1 ve 5.2'de GDSC veri kümesinde bulunan 265 ilaç sırasıyla IC_{50} ve AUC değerlerinin tahmininde hangi modelin hangi ilaç için daha iyi tahminleri ürettiği gösterilmiştir. Buradaki 265 ilacın listesi EK1'de verilmiştir.

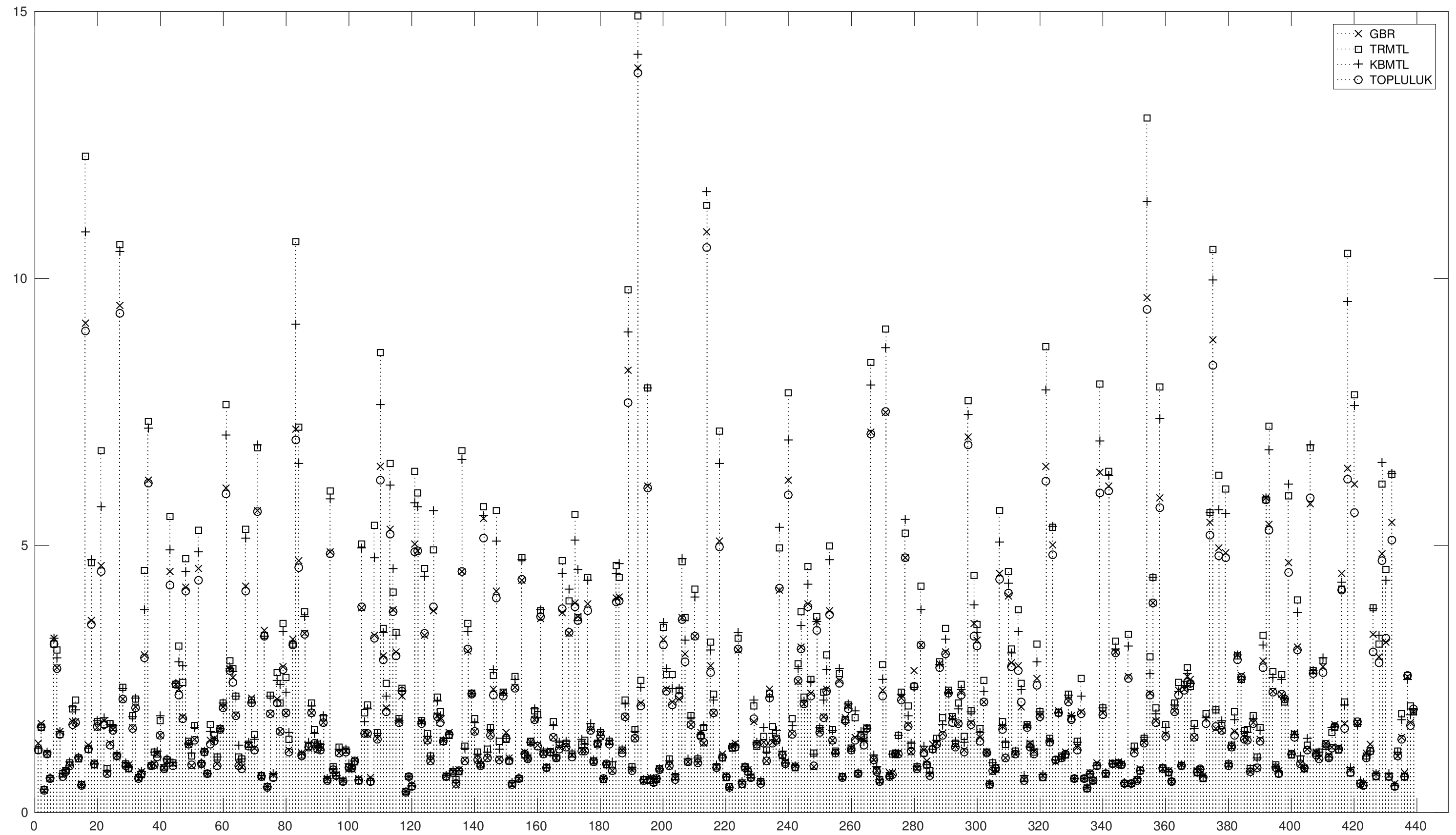
Şekil 5.3 ise CTRP veri kümesinde yer alan 439 ilaç için, modellerin her bir ilacın hücre hatları üzerindeki aktivite tahmini hatasını MSE cinsinden yansıtmaktadır. Buradaki ilaçların listesi de EK2'de verilmiştir.



Şekil 5.1: Topluluk yöntemi kullanılarak tahmin edilen GDSC (IC_{50}) verisi için ilaçların bireysel karşılaştırılması



Şekil 5.2: Topluluk yöntemi kullanılarak tahmin edilen GDSC (AUC) verisi için ilaçların bireysel karşılaştırılması



Şekil 5.3: Topluluk yöntemi kullanılarak tahmin edilen CTRP (*AUC*) verisi için ilaçların bireysel karşılaştırılması

Çizelge 5.5: GDSC veri kümesi için eksik değerlerin tahmini

İlaç	Hücre hattı	IC50
[3] Bortezomib	SW756	-7.50
[32]Docetaxel	HSC-3	-6.83
Epothilone B	IOSE-397	-6.07
GSK2126458	RCH-ACV	-5.86
AICAR	A673	-5.68
SN-38	SUP-B15	-5.59
YM155	OCI-LY7	-5.31
Vinorelbine	RCH-ACV	-4.83
Thapsigargin	IOSE-397	-4.79
PD-0325901	G-MEL	-4.62
Ispinesib Mesylate	RCH-ACV	-4.42
Paclitaxel	BICR78	-4.37
Gemcitabine	RCH-ACV	-4.28
Elesclomol	GAMG	-4.23
AP-24534	CML-T1	-4.21
LAQ824	RCH-ACV	-3.84
AUY922	IOSE-397	-3.79
Methotrexate	RS4-11	-3.66
Temsirolimus	SU-DHL-10	-3.65
GW843682X	BICR78	-3.64

5.5 Topluluk Modeli İçin Yeni Aktivite Tahminleri

Tüm veri kullanılarak topluluk modeli eğitildikten sonra, ilaçların hücre hatları üzerinde etkili olup olmadığının tanımlanması için, henüz deneylenmemiş ilaç-hücre hattı çiftlerinin değerleri tahmin edilmiştir. Oluşturulan modelin tahmini sonucunda en aktif olacağı düşünülen ilaçlar ve karşılık gelen hücre hatları Çizelge 5.5 ve Çizelge 5.6'de verilmiştir.

Tasarlanan model ile bilgisayar ortamında hangi hücre hatlarının hangi ilaçlara hassas olduğu ortaya çıkartılmıştır. Örneğin SW756 hücre hattı Bortezomib ilacına oldukça hassastır. Ayrıca bir hücre hattı üzerinde hangi ilacın daha aktif olacağı da tasarlanan model ile belirlenebilmektedir. Örnek olarak IOSE-397 hücre hattı üzerinde Epothilone B ilacının Thapsigargin ilacından daha etkili olduğu Çizelge 5.5 'den anlaşılmaktadır.

Uygulanmış canlı içi (in vivo) deneyler, GDSC verisi için oluşturulan modelin Çizelge 5.5'de verilen sonuçlarını desteklemektedir. Örneğin [3] tarafından yapılan bir çalışma, deneylenen tüm ilaçlar arasında Bortezomib ilacının SW756 hücre hattı üzerinde daha etkili olduğunu ortaya çıkarmıştır. Çalışmaya göre; Bortezomib ilacının eeyarestatin ile düşük derişimdeki birleşimi, SW756 kanser hücrelerinin klonal büyümelerini etkili bir biçimde baskı altına almıştır. Aynı zamanda başka bir çalışmada[32] da Docetaxel'in HSC-3 hücrelerindeki iyileştirici etkinliği geliştirmek için ester bağıyla polimer bloğuna konjuge edilmiştir ve Docetaxel ilacının tümörün gelişmesini kontrol ettiği gözlenmiştir.

Çizelge 5.6: CTRP veri kümesi için eksik değerlerin tahmini

İlaç	**Hücre hattı**	**AUC**
BRD-K13662825	D283MED	0.05
BRD-K27624156	TE14	1.69
[9] BRD-A05821830	NCIH226	1.78
BRD-A28746609	NCIH226	2.05
BRD-K02130563	AMO1	2.11
BRD-K82109576	NCIH1793	2.50
BRD-K92428232	NCIH1793	3.00
BRD-K23547378	NCIH1793	3.08
BRD-K76674262	NCIH1793	3.16
BRD-K62358710	SKNBE2	3.19
[6] BRD-K15108141	HT29	3.23
BRD-K37764012	D283MED	3.25
BRD-K64890080	NCIH1793	3.32
BRD-K35708212	AMO1	3.746
BRD-K92093830	NCIH1793	3.74
BRD-A35588707	ML1	3.78
BRD-K12343256	L363	3.79
BRD-K34022604	HEC1B	4.17
BRD-K58435339	VMCUB1	4.29
BRD-K29968218	5637	4.40

Çizelge 5.7: İmza benzerlik tabanlı regülarizasyonlu çoklu-iş öğrenme için GDSC veri kümesi (IC_{50}) sonuçları

	AMSE	**WAMSE**	**NDPB**
Lasso	1.97	1.94	30 / 133
İBTRÇÖ	**1.89**	**1.87**	**103** / 133

CTRP verisi kullanarak oluşturulan tahminler de benzer şekilde literatürdeki çalışmalar tarafından desteklenmektedir. Örneğin [9], LOR-253 ilacının BRD-A05821830 (Docetaxel) veya BRD-A28746609 (Paclitaxel) ile birlikte NCIH-256 üzerindeki sıralı ve eşzamanlı deneylerinde sinerji gözlenmiştir. Ayrıca [6] çalışmasında da BRD-K15108141 (gemcitabine) ilacının HT29 hücre hattı üzerinde etkili olduğu onaylanmaktadır.

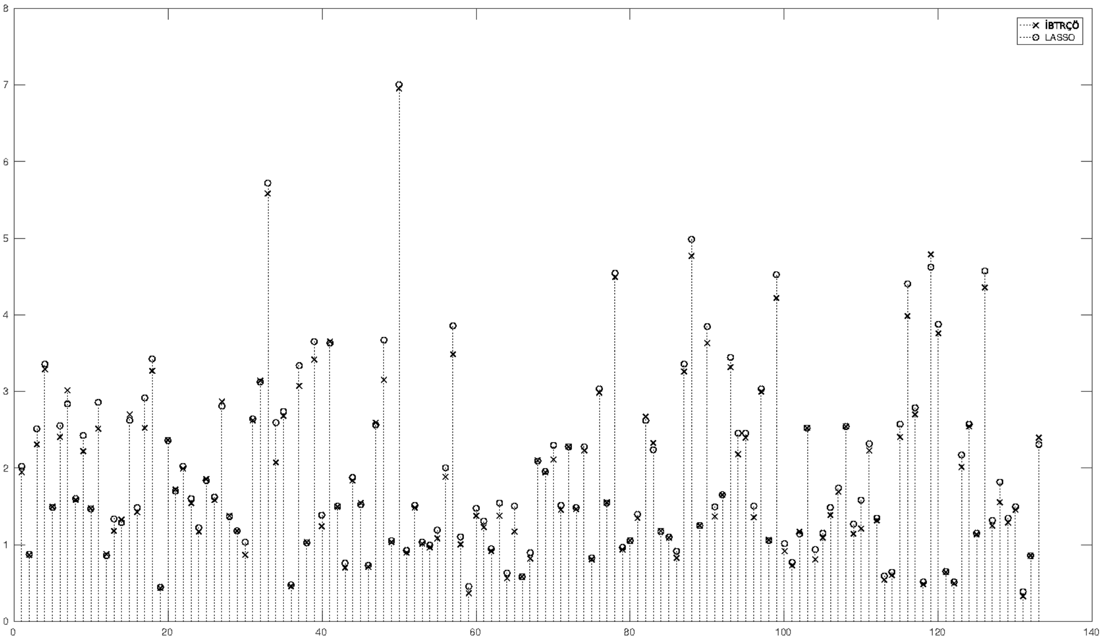
5.6 İBTRÇÖ Modeli İçin Çapraz Doğrulama

Oluşturulan modelin performansının ölçülmesi için, İBTRÇÖ modeli imzalar arasındaki benzerliğin göz ardı edildiği başka bir çoklu-iş öğrenme modeli (Lasso) ile karşılaştırılmıştır. Topluluk modelinde olduğu gibi üç farklı ölçev kullanılarak yapılan bu karşılaştırma sonucunda imza benzerlik tabanlı regülarizasyonlu çoklu-iş öğrenme modelinin kıyasla daha iyi performans gösterdiği her iki veri kümesi üzerinde de gösterilmiştir(Çizelge 5.7 ve Çizelge 5.8).

İlaçların ortalama karesel hatalarının(5.2) ortalaması (AMSE) (5.3), ilaçların ortalama

Çizelge 5.8: İmza benzerlik tabanlı regülarizasyonlu çoklu-iş öğrenme için CTRP veri kümesi (AUC) sonuçları

	AMSE	**WAMSE**	**NDPB**
Lasso	2.49	2.54	135 / 286
İBTRÇÖ	**2.47**	**2.53**	**151** / 286



Şekil 5.4: GDSC (IC_{50}) için ilaçların bireysel karşılaştırılması

karesel hatalarının ağırlıklı ortalaması (WAMSE)(5.4) ve her tahmin edicinin en iyi olarak tahmin ettiği ilaç sayısı (NDPB) kullanarak yapılan karşılaştırmada GDSC verisi için daha açık bir üstünlükten söz edilebilir. CTRP için ise hesaplanan sonuçlarda GDSC verisindeki kadar belirgin bir üstünlük görülmese de, bu veri kümesi için de sonuçlar olumludur.

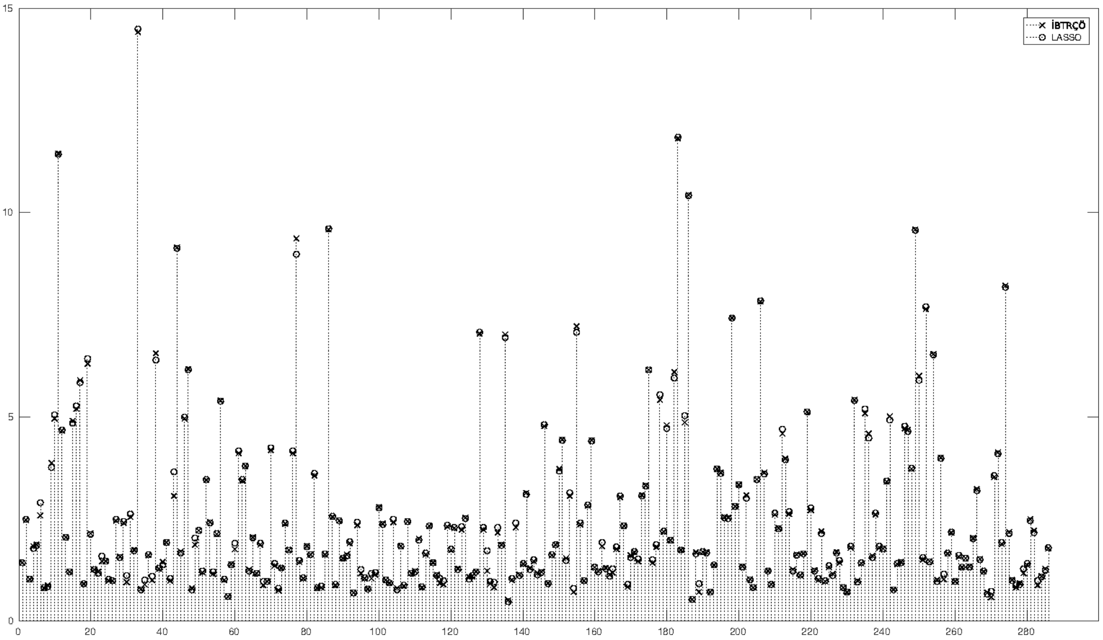
GDSC veri kümesindeki 133 ilaç için ayrı ayrı (Şekil 5.4), CTRP veri kümesindeki 286 ilaç için ayrı ayrı (Şekil 5.5) ilaçların MSE değerlerinin karşılaştırıldığı tablolar ile de bireysel olarak ilaçların tahmin hataları gösterilmiştir.

Ayrıca CTRP veri kümesinde daha az sayıdaki ilaçlar için de bu deneyler tekrarlanmıştır. 286 ilaç içerisinde; en fazla sayıda hücre hattı ile deneylenen 40 ilaç (Çizelge 5.9), en fazla sayıda deneye sahip 40 ilaç (Çizelge 5.10) ile ayrı ayrı yapılan deneylerdeki sonuçlar ile yine oluşturulan modelin katkısı gösterilmiştir.

CTRP verisi için listede bulunan 481 ilaçtan 286'sı için imza oluşturulmuştur. LINCS

Çizelge 5.9: CTRP veri kümesinde(AUC) en fazla sayıda hücre hattı ile deneylenen 40 ilaç için sonuçlar

	AMSE	**WAMSE**	**NDPB**
Lasso	2.62	2.61	14 / 40
İBTRÇÖ	**2.59**	**2.59**	**26** / 40



Şekil 5.5: CTRP (AUC) için ilaçların bireysel karşılaştırılması

Çizelge 5.10: CTRP veri kümesinde(AUC) en fazla sayıda deneye sahip 40 ilaç için sonuçlar

	AMSE	**WAMSE**	**NDPB**
Lasso	3.13	3.18	20 / 40
İBTRÇÖ	**3.11**	3.18	20 / 40

veri kümesine bakıldığında bu ilaçlardan bazılarının çok az sayıda hücre hattı ile deneylendiği gözlenmiştir. Bunun için, imzaların daha fazla deneylenmiş olmasının daha sağlam imza oluşturulması için gerekli bilgiyi vereceği düşüncesiyle, deney sayısı olarak ve deneylenen hücre sayısı olarak belli bir değerin üstündeki ilaçlar çalışmada kullanılmıştır. Bu çalışmanın sonucunda özellikle deneylenen hücre sayısına göre belirlenen ilaçlarda sonuçların olumlu olduğu gözlenmiştir. Bu da ileride LINCS veri kümesindeki deney sayısı arttığında tüm ilaçlar için daha sağlam modeller oluşturulabileceğinin göstergesidir.

6. SONUÇ

Kanser tedavisi için paylaşılan veri kümelerindeki gelişmenin etkisiyle, ilaç hassasiyet analizinde bilgisayar üzerinde yapılan (*in slico*) modellerin kullanımı önem kazanmıştır. Bu durum hem analizlerin daha hızlı bir şekilde yapılmasını sağlamaktadır hem de araştırmacıları ve hastaları oldukça yüksek bir maliyetten kurtarmaktadır.

Bu çalışmanın ilk kısmında ilaç tepkilerini tahmin etmek için üç farklı öğrenme modelini; gradyan destekli regresyon, iz-norm regülarizasyonlu çoklu-iş öğrenme ve çekirdekli bayes çoklu-iş öğrenme, yığıtlı genelleme yöntemiyle birleştiren bir topluluk modeli tasarlanmıştır. Oluşturulan modelin geçerliliğinin onaylanması için ilaç-hücre hattı çiftlerinin aktivite verilerini içeren iki büyük veri kümesi kullanılmıştır. Her iki veri kümesi üzerindeki çapraz doğrulama sonuçları, oluşturulan modelin diğerlerini performans olarak geride bıraktığını göstermektedir. Bu sonuçlar ışığında, orijinal veri setlerinde bulunmayan çiftler için yeni tahminler yapılmıştır. Oluşturulan modelin tahminlerindeki iyileşme, birleştirilecek uygun modelleri seçmekten ve bu modelleri birleştirme yönteminden kaynaklanmaktadır.

Bu yöntem üzerinde çalışılması planlanan bir kaç eklenti daha vardır. İlk olarak, tasarlanan topluluk modeli, farklı öğrenme modelleri seçilerek veya seçilen modelleri birleştirmek için başka yollar kullanılarak genişletilebilir. Belirtildiği gibi, bu seçimler topluluk öğrenme modellerinde kilit rol oynamaktadır.

Bunun yanı sıra, farklı adlandırmalardan kaynaklanan ilaçların veya hücre hatlarının uyuşma problemlerini ortadan kaldırdıktan sonra (ör. Hücre hatları eş anlamlıdır ve farklı verisetlerinde farklı olarak adlandırılmıştır.) model, farklı veri setlerini birleştirerek her bir ilaç için daha dengeli veri setleri kullanarak eğitilebilir. Ayrıca, seçilen ve kullanılan öznitelikler, öğrenme modelleri için önemlidir. İlaç tepkisi tahmini için geleneksel yol, gen ifade verilerini kullanılmasıdır ancak bu özellikler ilaçların kimyasal bilgisinin ve diğer genomik özelliklerinin entegrasyonu ile genişletilebilir.

Tez çalışmasında aktarılan diğer yöntem ise ilaçların imzalarının elde edilmesi, imzalardan benzerlik ilişkisinin ortaya konması ve bu benzerlik ilişkilerinden yararlanan bir çoklu-iş öğrenme modelinin oluşturulmasını içerir. İmza benzerlik tabanlı çoklu-iş öğrenme yöntemi olarak adlandırılan bu yöntemin ilaç aktivite tahmini için iki farklı veri kümesiyle yapılan çapraz değerlendirmelerin sonuçlarına bakıldığında, oluşturulan modelin benzerlik ilişkisinin verilmediği modelden daha iyi sonuçlar ortaya koyduğu gözlenmiştir.

LINCS veri kümesinin ve yardımcı bazı tabloların geliştirilmesiyle bu modelin daha da güçleneceği düşünülmektedir. İlaçlar farklı veri kümelerinde farklı kimlikler ile yer aldığı için ilaç imzalarının oluşturulması aşamasında problemlerle karşılaşılmıştır. GDSC veri kümesinde bulunan ilaçların yarısından bile azı için imza üretilebilmiştir. İleride bu eşleştirme işlemlerinin daha rahat şekilde yapılabilmesi için gerekli olan arama tablolarının oluşturulmasıyla, daha fazla sayıda ilaç için imza oluşturulacağı düşünülmektedir. Ayrıca LINCS tarafından, veri kümesinde bulunan deneylerin de geliştirilmesiyle bu imzaların güvenilirliği de artacaktır.

GDSC ve CTRP veri kümelerinde yüzlerce hücre hattı bulunurken, LINCS’te üzerinde deney yapılan hücre hattı sayısı şuan 76’dır. İleride diğer hücre hatlarıyla da yapılcak deneylerin eklenmesi daha sağlam bir model oluşturulmasına katkı sağlayacaktır. Bu projenin devamında, ilaç aktivite imzası oluşturma işlemine benzer olarak, hücre aktivite imzasının da oluşturulması ve modelde kullanılması da amaçlanmaktadır. Yapılması planlanan bir diğer çalışma da verilen modellerin bir arada kullanılmasıdır. İmza benzerlik tabanlı regülarizasyonlu çoklu-iş öğrenme modeli, topluluk yöntemi için temel alınan modellerden biri olarak düşünülebilir.

Özetle, bu çalışmada ilaç aktivite tahmini için iki farklı yöntem geliştirilmiştir. İlk çalışma olan topluluk yönteminde, temel alınan modeller yığıtlı genelleme ile birleştirilmiştir. İkinci çalışmada ise ilaç imzaları ile hesaplanan benzerlik ilişkisi kullanılarak regülüze edilmiş çoklu-iş öğrenme modeli oluşturulmuştur. Yapılan değerlendirmelerde her iki modelin de ilaç aktivite tahmini için literatüre önemli katkılar yaptığı söylenebilir. Tezin ilk kısmındaki topluluk yönteminin ele alındığı bir makale, uluslararası bir konferansta [3] yayımlanmıştır. Ayrıca her iki çalışma da TÜBİTAK (Proje No: 115E274) tarafından desteklenmektedir.

---

[3] 8th International Conference on Knowledge Discovery and Information Retrieval

KAYNAKLAR

[1] **Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., ve diğ.** The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature 483*, 7391 (3 2012), 603–307.

[2] **Bellmunt, J., Werner, L., Bamias, A., Fay, A. P., ve diğ.** HER2 as a target in invasive urothelial carcinoma. *Cancer Medicine 4*, 6 (6 2015), 844–852.

[3] **Brem, G. J., Mylonas, I., ve Brüning, A.** Eeyarestatin causes cervical cancer cell sensitization to bortezomib treatment by augmenting ER stress and CHOP expression. *Gynecologic Oncology 128* (2013), 383–390.

[4] **Byers, L. A., Diao, L., ve diğ.** An Epithelial-Mesenchymal Transition Gene Signature Predicts Resistance to EGFR and PI3K Inhibitors and Identifies Axl as a Therapeutic Target for Overcoming EGFR Inhibitor Resistance. *Clinical Cancer Research 19*, 1 (1 2013), 279–290.

[5] **Caruana, R.** Multitask Learning. In *Learning to Learn*. Springer US, Boston, MA, 1998, pp. 95–133.

[6] **Chan, G. K. Y., Kleinheinz, T. L., ve diğ.** A simple high-content cell cycle assay reveals frequent discrepancies between cell number and ATP and MTS proliferation assays. *PloS one 8*, 5 (2013), e63583.

[7] **Cortés-Ciriano, I., van Westen, G. J. P., Bouvier, G., ve diğ.** Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics 32*, 1 (9 2015), btv529.

[8] **Costello, J. C., Heiser, L. M., Georgii, E., ve diğ.** A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology 32*, 12 (2014), 1202.

[9] **Cukier, H., Peralta, R., Jin, H., ve diğ.** Preclinical dose scheduling studies of lor-253, a novel anticancer drug, in combination with chemotherapeutics in lung and colon cancers. In *Annual Meeting of the American Association for Cancer Research; Chicago, IL. Philadelphia (PA): AACR; Cancer Res* (2012), vol. 72.

[10] **DeNardo, D. G., Brennan, D. J., Rexhepaj, E., ve diğ.** Leukocyte Complexity Predicts Breast Cancer Survival and Functionally Regulates Response to Chemotherapy. *Cancer Discovery 1*, 1 (2011).

[11] **Dietterich, T. G.** Ensemble Methods in Machine Learning. Springer Berlin Heidelberg, 2000, pp. 1–15.

[12] **Dong, Z., Zhang, N., Li, C., ve diğ.** Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC cancer 15*, 1 (2015), 489.

[13] **Duan, Q., Flynn, C., Niepel, M., Hafner, M., Muhlich, J. L., Fernandez, N. F., Rouillard, A. D., Tan, C. M., Chen, E. Y., Golub, T. R., ve diğ.** Lincs canvas browser: interactive web app to query, browse and interrogate lincs l1000 gene expression signatures. *Nucleic acids research* (2014), gku476.

[14] **Falgreen, S., Dybkær, K., Young, K. H., Xu-Monette, Z. Y., El-Galaly, T. C., Laursen, M. B., Bødker, J. S., Kjeldsen, M. K., Schmitz, A., Nyegaard, M., Johnsen, H. E., ve Bøgsted, M.** Predicting response to multidrug regimens in cancer patients using cell line experiments and regularised regression models. *BMC Cancer 15*, 1 (12 2015), 235.

[15] **Fersini, E., Messina, E., Archetti, F., ve diğ.** A p-Median approach for predicting drug response in tumour cells. *BMC Bioinformatics 15*, 1 (12 2014), 353.

[16] **Friedman, J. H.** Stochastic gradient boosting. *Computational Statistics and Data Analysis 38*, 4 (2002), 367–378.

[17] **Gholami, A., Hahne, H., Wu, Z., Auer, F., Meng, C., Wilhelm, M., ve Kuster, B.** Global Proteome Analysis of the NCI-60 Cell Line Panel. *Cell Reports 4*, 3 (8 2013), 609–620.

[18] **Gonen, M., ve Margolin, A. A.** Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning. *Bioinformatics 30*, 17 (9 2014), i556–i563.

[19] **Jackson, S. E., ve Chester, J. D.** Personalised cancer medicine. *International Journal of Cancer 137*, 2 (7 2015), 262–266.

[20] **Jang, I. S., Neto, E. C., Guinney, J., Friend, S. H., ve Margolin, A. a.** Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2014), 63–74.

[21] **Ji, S., ve Ye, J.** An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th annual international conference on machine learning* (2009), ACM, pp. 457–464.

[22] **Menden, M. P., Iorio, F., Garnett, M., ve diğ.** Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS ONE 8*, 4 (4 2013), e61318.

[23] **National Cancer Database**. American College of Surgeons Commission on Cancer, 2011 Data Submission. *American College of Surgeons* (2013).

[24] **Neto, E. C., Jang, I. S., Friend, S. H., ve Margolin, A. A.** The Stream algorithm: computationally efficient ridge-regression via Bayesian model averaging, and applications to pharmacogenomic prediction of cancer cell line sensitivity. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2014), 27–38.

[25] **Nicolau, M., Levine, A. J., ve Carlsson, G.** Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences 108*, 17 (4 2011), 7265–7270.

[26] **Qin, Y., Chen, M., ve diğ.** A network flow-based method to predict anticancer drug sensitivity. *PLoS ONE 10*, 5 (2015), 1–14.

[27] **Rappaport, N., Nativ, N., Stelzer, G., ve diğ.** MalaCards: an integrated compendium for diseases and their annotation. *Database : the journal of biological databases and curation 2013* (2013), bat018.

[28] **Rees, M. G., Seashore-Ludlow, B., Cheah, J. H., ve diğ.** Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nature Chemical Biology 12*, 2 (12 2015), 109–116.

[29] **Riddick, G., Song, H., Ahn, S., ve diğ.** Predicting in vitro drug sensitivity using Random Forests. 220–22410.

[30] **Seashore-Ludlow, B., Rees, M. G., Cheah, J. H., ve diğ.** Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer discovery 5*, 11 (11 2015), 1210–23.

[31] **Sewell, M.** Ensemble learning. *RN 11*, 02.

[32] **Shi, L., Song, X.-B., Wang, Y., ve diğ.** Docetaxel-conjugated monomethoxy-poly(ethylene glycol)-b-poly(lactide) (mPEG-PLA) polymeric micelles to enhance the therapeutic efficacy in oral squamous cell carcinoma. *RSC Adv. 6*, 49 (2016), 42819–42826.

[33] **Shoemaker, R. H.** The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer 6*, 10 (10 2006), 813–823.

[34] **Siegel, R. L., Miller, K. D., ve Jemal, A.** Cancer statistics, 2016. *CA: A Cancer Journal for Clinicians 66*, 1 (1 2016), 7–30.

[35] **Tan, M.** Prediction of anti-cancer drug response by kernelized multi-task learning. *Artificial Intelligence in Medicine 73* (2016), 70–77.

[36] **Wan, Q., ve Pal, R.** An ensemble based top performing approach for nci-dream drug sensitivity prediction challenge. *PloS one 9*, 6 (2014), e101183.

[37] **Wolpert, D. H.** Stacked generalization. *Neural Networks 5*, 2 (1992), 241–259.

[38] **Wu, C., MacLeod, I., ve Su, A. I.** Biogps and mygene. info: organizing online, gene-centric information. *Nucleic acids research* (2012), gks1114.

[39] **Yang, W., Soares, J., Greninger, P., ve diğ.** Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research 41*, D1 (1 2013), D955–D961.

[40] **Yuan, H., Paskov, I., Paskov, H., González, A. J., ve Leslie, C. S.** Multitask learning improves prediction of cancer drug sensitivity. *Scientific reports 6* (8 2016), 31619.

[41] **Zhang, N., Wang, H., Fang, Y., ve diğ.** Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model. *PLoS computational biology 11*, 9 (2015), e1004498.

[42] **Zhao, Z., Fu, G., Liu, S., ve diğ.** Drug activity prediction using multiple-instance learning via joint instance and feature selection. *BMC Bioinformatics 14 Suppl 1*, Suppl 14 (2013), S16.

[43] **Zhou, J.** Multi-task learning in crisis event classification. Tech. rep., Tech. Rep., http://www. public. asu. edu/jzhou29.

[44] **Zhou, J., Chen, J., ve Ye, J.** User's Manual MALSAR: Multi-tAsk Learning via StructurAl Regularization. *Arizona State University* (2012).

EKLER

EK 1 : Şekil 5.1 ve 5.2 için İlaç Listesi

EK 2 : Şekil 5.3 için İlaç Listesi

EK 3 : Şekil 5.4 için İlaç Listesi

EK 4 : Şekil 5.5 için İlaç Listesi

EK 1

1. Erlotinib
2. Rapamycin
3. Sunitinib
4. PHA-665752
5. MG-132
6. Paclitaxel
7. Cyclopamine
8. AZ628
9. Sorafenib
10. VX-680
11. Imatinib
12. TAE684
13. Crizotinib
14. Saracatinib
15. S-Trityl-L-cysteine
16. Z-LLNle-CHO
17. Dasatinib
18. GNF-2
19. CGP-60474
20. CGP-082996
21. A-770041
22. WH-4-023
23. WZ-1-84
24. BI-2536
25. BMS-536924
26. BMS-509744
27. CMK
28. Pyrimethamine
29. JW-7-52-1
30. A-443654
31. GW843682X
32. MS-275
33. Parthenolide
34. KIN001-135
35. TGX221
36. Bortezomib
37. XMD8-85
38. Roscovitine
39. Salubrinal
40. Lapatinib
41. GSK269962A
42. Doxorubicin
43. Etoposide
44. Gemcitabine
45. Mitomycin C
46. Vinorelbine
47. NSC-87877
48. Bicalutamide
49. QS11
50. CP466722
51. Midostaurin
52. CHIR-99021
53. AP-24534
54. AZD6482
55. JNK-9L
56. PF-562271
57. HG-6-64-1
58. JQ1
59. JQ12
60. DMOG
61. FTI-277
62. OSU-03012
63. Shikonin
64. AKT inhibitor VIII
65. Embelin
66. FH535
67. PAC-1
68. IPA-3
69. GSK-650394
70. BAY 61-3606
71. 5-Fluorouracil
72. Thapsigargin
73. Obatoclax Mesylate
74. BMS-754807
75. Lisitinib
76. Bexarotene
77. Bleomycin
78. LFM-A13
79. GW-2580
80. AUY922
81. Phenformin
82. Bryostatin 1
83. Pazopanib
84. LAQ824
85. Epothilone B
86. GSK1904529A
87. BMS345541
88. Tipifarnib
89. BMS-708163
90. Ruxolitinib
91. AS601245
92. Ispinesib Mesylate
93. TL-2-105
94. AT-7519
95. TAK-715
96. BX-912
97. ZSTK474
98. AS605240
99. Genentech Cpd 10
100. GSK1070916
101. KIN001-102
102. LY317615
103. GSK429286A
104. FMK
105. QL-XII-47
106. CAL-101
107. UNC0638
108. XL-184
109. WZ3105
110. XMD14-99
111. AC220
112. CP724714
113. JW-7-24-1
114. NPK76-II-72-1
115. STF-62247
116. NG-25
117. TL-1-85
118. VX-11e
119. FR-180204
120. Tubastatin A
121. Zibotentan
122. YM155
123. NSC-207895
124. VNLG/124
125. AR-42
126. CUDC-101
127. Belinostat
128. I-BET-762
129. CAY10603
130. Linifanib
131. BIX02189
132. CH5424802
133. EKB-569
134. GSK2126458
135. KIN001-236
136. KIN001-244
137. KIN001-055
138. KIN001-260
139. KIN001-266
140. Masitinib
141. MP470
142. MPS-1-IN-1
143. BHG712
144. OSI-930
145. OSI-027
146. CX-5461
147. PHA-793887
148. PI-103
149. PIK-93
150. SB52334
151. TPCA-1
152. TG101348
153. Foretinib
154. Y-39983
155. YM201636
156. Tivozanib
157. GSK690693
158. SNX-2112
159. QL-XI-92
160. XMD13-2
161. QL-X-138
162. XMD15-27
163. T0901317
164. EX-527
165. THZ-2-49
166. KIN001-270
167. THZ-2-102-1
168. AICAR
169. Camptothecin
170. Vinblastine
171. Cisplatin
172. Cytarabine
173. Docetaxel
174. Methotrexate
175. ATRA
176. Gefitinib
177. Navitoclax
178. Vorinostat
179. Nilotinib
180. RDEA119
181. CI-1040
182. Temsirolimus
183. Olaparib
184. Veliparib
185. Bosutinib
186. Lenalidomide
187. Axitinib
188. AZD7762
189. GW 441756
190. CEP-701
191. SB 216763
192. 17-AAG
193. VX-702
194. AMG-706
195. KU-55933
196. Elesclomol
197. Afatinib
198. GDC0449
199. PLX4720
200. BX-795
201. NU-7441
202. SL 0101-1
203. BIRB 0796
204. JNK Inhibitor VIII
205. 681640
206. Nutlin-3a (-)
207. PD-173074
208. ZM-447439
209. RO-3306
210. MK-2206
211. PD-0332991
212. BEZ235
213. GDC0941
214. AZD8055
215. PD-0325901
216. SB590885
217. selumetinib
218. AZD6482
219. CCT007093
220. EHT 1864
221. BMS-708163
222. BMS-536924
223. Cetuximab
224. PF-4708671
225. JNJ-26854165
226. HG-5-113-01
227. HG-5-88-01
228. TW 37
229. XMD11-85h
230. ZG-10
231. XMD8-92
232. QL-VIII-58
233. CCT018159
234. AG-014699
235. GSK269962A
236. SB 505124
237. Tamoxifen
238. QL-XII-61
239. JQ1
240. PFI-1
241. IOX2
242. UNC0638
243. YK 4-279
244. CHIR-99021
245. (5Z)-7-Oxozeaenol
246. piperlongumine
247. FK866
248. Talazoparib
249. rTRAIL
250. UNC1215
251. SGC0946
252. XAV939
253. PLX4720
254. Trametinib
255. Dabrafenib
256. Temozolomide
257. Afatinib
258. Bleomycin (50 uM)
259. SN-38
260. Olaparib
261. selumetinib
262. Bicalutamide
263. RDEA119
264. GDC0941
265. MLN4924

EK 2

1. BRD-K46556387
2. BRD-K86574132
3. BRD-K35716340
4. BRD-K89692698
5. BRD-A18763547
6. BRD-K89329876
7. BRD-K64634304
8. BRD-K45401373
9. BRD-K15563106
10. BRD-A93255169
11. BRD-K19295594
12. BRD-K29458283
13. BRD-K24844714
14. BRD-K03406345
15. BRD-K89732114
16. BRD-A28746609
17. BRD-K93754473
18. BRD-A35588707
19. BRD-K50128260
20. BRD-K22134346
21. BRD-K02407574
22. BRD-K13032584
23. BRD-K74148702
24. BRD-K13044802
25. BRD-K55591206
26. BRD-K43149758
27. BRD-K59456551
28. BRD-A70155556
29. BRD-K12994359
30. BRD-A09722536
31. BRD-K35520305
32. BRD-K35960502
33. BRD-K19352500
34. BRD-A67097164
35. BRD-K92093830
36. BRD-K35708212
37. BRD-K05649647
38. BRD-K24132293
39. BRD-K26531177
40. BRD-A80213327
41. BRD-K01567962
42. BRD-K08608874
43. BRD-K55696337
44. BRD-K71726959
45. BRD-A02481876
46. BRD-K37798499
47. BRD-K15318909
48. BRD-K81473043
49. BRD-K26664453
50. BRD-K68143200
51. BRD-K78599730
52. BRD-K59670716
53. BRD-K69608737
54. BRD-K66874953
55. BRD-K67298865
56. BRD-K61323504
57. BRD-K26603252
58. BRD-K52075040
59. BRD-K73397362
60. BRD-K04923131
61. BRD-K49328571
62. BRD-K64052750
63. BRD-K70401845
64. BRD-K94991378
65. BRD-K26818574
66. BRD-K52313696
67. BRD-K02130563
68. BRD-K75430629
69. BRD-A38030642
70. BRD-K81418486
71. BRD-K84937637
72. BRD-K25311561
73. BRD-K06854232
74. BRD-K05653692
75. BRD-K71935468
76. BRD-K60230970
77. BRD-K77908580
78. BRD-K17743125
79. BRD-K64606589
80. BRD-K04466929
81. BRD-A94377914
82. BRD-A42556028
83. BRD-K82109576
84. BRD-K33106058
85. BRD-A34817987
86. BRD-A25004090
87. BRD-K38985961
88. BRD-K11740178
89. BRD-A91658086
90. BRD-K92856060
91. BRD-K10882151
92. BRD-K45681478
93. BRD-A83255679
94. BRD-A36275421
95. BRD-K05392795
96. BRD-K29711668
97. BRD-K20215950
98. BRD-K64497429
99. BRD-K59204667
100. BRD-K61128465
101. BRD-K44241590
102. BRD-A86956638
103. BRD-K45841694
104. BRD-A35108200
105. BRD-A40802033
106. BRD-K92723993
107. BRD-K92041145
108. BRD-K79254416
109. BRD-K16189898
110. BRD-K64890080
111. BRD-K61166597
112. BRD-K29905972
113. BRD-K57080016
114. BRD-K12184916
115. BRD-K90382497
116. BRD-K77625799
117. BRD-K23984367
118. BRD-K59962020
119. BRD-A94153989
120. BRD-K90370028
121. BRD-K71281111
122. BRD-K82746043
123. BRD-K92441787
124. BRD-K81528515
125. BRD-K42828737
126. BRD-K17068645
127. BRD-K76674262
128. BRD-K28907958
129. BRD-K17140735
130. BRD-K06593056
131. BRD-K20755323
132. BRD-K88544581
133. BRD-A36318220
134. BRD-K10705233
135. BRD-K93176058
136. BRD-K00317371
137. BRD-K37865504
138. BRD-K47150025
139. BRD-K25737009
140. BRD-K51575138
141. BRD-K37392901
142. BRD-K96799727
143. BRD-A36630025
144. BRD-K80183349
145. BRD-K66532283
146. BRD-K68065987
147. BRD-K05402890
148. BRD-K66453893
149. BRD-K11533227
150. BRD-K14844214
151. BRD-K62801835
152. BRD-K41597374
153. BRD-K63431240
154. BRD-K13999467
155. BRD-A31107743
156. BRD-K32536677
157. BRD-K87317732
158. BRD-K83213911
159. BRD-A13122391
160. BRD-A12230535
161. BRD-K07303502
162. BRD-K98404142
163. BRD-K09907482
164. BRD-K08417745
165. BRD-K53903639
166. BRD-K40853697
167. BRD-K88510285
168. BRD-K56301217
169. BRD-A01145011
170. BRD-K67868012
171. BRD-K00627859
172. BRD-K11172604
173. BRD-K81651477
174. BRD-K61829047
175. BRD-K12502280
176. BRD-K67844266
177. BRD-K52836380
178. BRD-K32330832
179. BRD-K80672993
180. BRD-K53855319
181. BRD-K61662457
182. BRD-K30064966
183. BRD-K89391146
184. BRD-K92991072
185. BRD-K98538768
186. BRD-K47335880
187. BRD-K10466330
188. BRD-K34581968
189. BRD-K15108141
190. BRD-K19894101
191. BRD-K02113016
192. BRD-A81541225
193. BRD-K17610631
194. BRD-K02492147
195. BRD-K25987073
196. BRD-K93095519
197. BRD-K09587429
198. BRD-K44741158
199. BRD-K00088062
200. BRD-K62965247
201. BRD-K63195589
202. BRD-K67112618
203. BRD-K38264551
204. BRD-A22997170
205. BRD-K20285085
206. BRD-K76703230
207. BRD-K13049116
208. BRD-K04623885
209. BRD-K92241597
210. BRD-K85606544
211. BRD-K73261812
212. BRD-K68548958
213. BRD-K60219430
214. BRD-K58550667
215. BRD-K54256913
216. BRD-K53414658
217. BRD-K64642496
218. BRD-K23547378
219. BRD-K71781559
220. BRD-K04800985
221. BRD-K19103580
222. BRD-K64610608
223. BRD-K02251932
224. BRD-K55116708
225. BRD-K16147474
226. BRD-K33199242
227. BRD-K52037352
228. BRD-K37390332
229. BRD-K89162000
230. BRD-K55478147
231. BRD-K17349619
232. BRD-K39706510
233. BRD-K45748132
234. BRD-K49456190
235. BRD-A68631409
236. BRD-K86856088
237. BRD-A28105619
238. BRD-K50799972
239. BRD-K17060750
240. BRD-K63923597
241. BRD-K99749624
242. BRD-K87142802
243. BRD-K19540840
244. BRD-K66175015
245. BRD-K86930074
246. BRD-K50168500
247. BRD-K15600710
248. BRD-K71035033
249. BRD-K63068307
250. BRD-K28428262
251. BRD-K67578145
252. BRD-K50140147
253. BRD-K43389698
254. BRD-K16478699
255. BRD-A41692738
256. BRD-K67566344
257. BRD-K64785675
258. BRD-K06750613
259. BRD-K02965346
260. BRD-K88742110
261. BRD-K01436366
262. BRD-K22828899
263. BRD-K76964878
264. BRD-K44847641
265. BRD-K36852164
266. BRD-K23853216
267. BRD-K69982010
268. BRD-K24376488
269. BRD-K36739687
270. BRD-K35498412
271. BRD-K01877528
272. BRD-K35723520
273. BRD-A56592690
274. BRD-K09485525
275. BRD-K83336168
276. BRD-K28360340
277. BRD-K69932463
278. BRD-K94145482
279. BRD-K01121114
280. BRD-A67788537
281. BRD-A63646118
282. BRD-K54606188
283. BRD-K51490254
284. BRD-K85133207
285. BRD-K13566078
286. BRD-K51781482
287. BRD-K47503321
288. BRD-K49075727
289. BRD-K78431006
290. BRD-K03449891

291. BRD-K16730910
292. BRD-K81728688
293. BRD-K25340465
294. BRD-K56277358
295. BRD-K39974922
296. BRD-A62182663
297. BRD-K83289131
298. BRD-K09499853
299. BRD-K06792661
300. BRD-K78659596
301. BRD-K69840642
302. BRD-K19687926
303. BRD-K29313308
304. BRD-K74236984
305. BRD-K19477839
306. BRD-K19796430
307. BRD-K75295174
308. BRD-K58772419
309. BRD-K53972329
310. BRD-A42083487
311. BRD-K52911425
312. BRD-K37720887
313. BRD-K64800655
314. BRD-K93918653
315. BRD-K88025533
316. BRD-K74514084
317. BRD-K72420232
318. BRD-K24690302
319. BRD-K62825658
320. BRD-K56343971
321. BRD-K48477130
322. BRD-K92428232
323. BRD-K28456706
324. BRD-K22477529
325. BRD-K93367411
326. BRD-K84964099
327. BRD-K96335988
328. BRD-A02303741
329. BRD-K81491172
330. BRD-A34462049
331. BRD-K79877282
332. BRD-K79239947
333. BRD-K31514534
334. BRD-K78667050
335. BRD-K81335284
336. BRD-K03391209
337. BRD-A76527075
338. BRD-K66430217
339. BRD-K27624156
340. BRD-K59146805
341. BRD-K31406481
342. BRD-K97651142
343. BRD-K93442924
344. BRD-K71487808
345. BRD-A71883111
346. BRD-K50145167
347. BRD-K60750172
348. BRD-K36363294
349. BRD-K27955832
350. BRD-K11593101
351. BRD-K99655327
352. BRD-K45935533
353. BRD-K35604418
354. BRD-K62358710
355. BRD-K08109215
356. BRD-K92960067
357. BRD-K15179513
358. BRD-K34022604
359. BRD-K75308783
360. BRD-K65928735
361. BRD-A82720512
362. BRD-K47981327
363. BRD-K59437938
364. BRD-A04287157
365. BRD-K04905989
366. BRD-A52530402
367. BRD-K28392481
368. BRD-A30032852
369. BRD-A12571627
370. BRD-K18850819
371. BRD-K65592642
372. BRD-A05715709
373. BRD-K90746395
374. BRD-K08177763
375. BRD-A05821830
376. BRD-K84466663
377. BRD-K55187425
378. BRD-K12040459
379. BRD-A57798112
380. BRD-K36016295
381. BRD-K74065929
382. BRD-K21718444
383. BRD-K93123848
384. BRD-K81458380
385. BRD-K87737963
386. BRD-K60866521
387. BRD-K43644456
388. BRD-K87124298
389. BRD-K02017404
390. BRD-K29395450
391. BRD-K86525559
392. BRD-K18589165
393. BRD-K29968218
394. BRD-A09890259
395. BRD-K78978711
396. BRD-K54640016
397. BRD-K51544265
398. BRD-M00053801
399. BRD-K71467466
400. BRD-K90570971
401. BRD-K14870255
402. BRD-K70301465
403. BRD-K16761703
404. BRD-K76894955
405. BRD-A59431241
406. BRD-K62391742
407. BRD-K24556407
408. BRD-K57261999
409. BRD-K52560704
410. BRD-K54997624
411. BRD-K49215523
412. BRD-K75664313
413. BRD-K64117221
414. BRD-K50501969
415. BRD-K42137908
416. BRD-K99006945
417. BRD-K33379087
418. BRD-K12343256
419. BRD-K22024824
420. BRD-K13662825
421. BRD-K43410529
422. BRD-K51831558
423. BRD-K87774949
424. BRD-K90860366
425. BRD-K42260513
426. BRD-K68682020
427. BRD-K84754008
428. BRD-A86708339
429. BRD-K37764012
430. BRD-K09951645
431. BRD-K55125219
432. BRD-K58435339
433. BRD-A68258842
434. BRD-K83607951
435. BRD-K34222889
436. BRD-K43428468
437. BRD-A15100685
438. BRD-K08589866
439. BRD-K02834582

EK 3

1. A-443654
2. AMG-706
3. AP-24534
4. AT-7519
5. ATRA
6. AUY922
7. AZ628
8. AZD6482
9. Afatinib
10. Axitinib
11. BI-2536
12. BIRB0796
13. BMS-509744
14. BMS-536924
15. BMS-754807
16. BX-795
17. BX-912
18. Belinostat
19. Bicalutamide
20. Bortezomib
21. Bosutinib
22. CEP-701
23. CGP-60474
24. CHIR-99021
25. CI-1040
26. CP466722
27. Camptothecin
28. Cisplatin
29. Crizotinib
30. Cyclopamine
31. Cytarabine
32. Dabrafenib
33. Dasatinib
34. Docetaxel
35. Doxorubicin
36. EX-527
37. Elesclomol
38. Embelin
39. Epothilone B
40. Erlotinib
41. Etoposide
42. FH535
43. FR-180204
44. Foretinib
45. GDC0941
46. GNF-2
47. GSK-650394
48. GW843682X
49. Gefitinib
50. Gemcitabine
51. HG-5-113-01
52. HG-5-88-01
53. Imatinib
54. JNJ-26854165
55. JNK-9L
56. JW-7-24-1
57. JW-7-52-1
58. KIN001-055
59. KIN001-135
60. KIN001-244
61. KIN001-266
62. KU-55933
63. LAQ824
64. LFM-A13
65. Lapatinib
66. Lenalidomide
67. Linifanib
68. MG-132
69. MK-2206
70. MS-275
71. Masitinib
72. Methotrexate
73. Midostaurin
74. NPK76-II-72-1
75. NU-7441
76. Navitoclax
77. Nilotinib
78. OSI-027
79. OSI-930
80. Olaparib
81. PAC-1
82. PD-0325901
83. PD-0332991
84. PD-173074
85. PF-562271
86. PHA-665752
87. PHA-793887
88. PI-103
89. PLX4720
90. Paclitaxel
91. Parthenolide
92. Pazopanib
93. Phenformin
94. Pyrimethamine
95. QL-X-138
96. QL-XI-92
97. QL-XII-47
98. RO-3306
99. Rapamycin
100. Roscovitine
101. Ruxolitinib
102. SB590885
103. SN-38
104. STF-62247
105. Salubrinal
106. Saracatinib
107. Sorafenib
108. Sunitinib
109. TAK-715
110. TGX221
111. TPCA-1
112. TW 37
113. Tamoxifen
114. Temozolomide
115. Temsirolimus
116. Thapsigargin
117. Tipifarnib
118. Tivozanib
119. Trametinib
120. VX-680
121. VX-702
122. Veliparib
123. Vinblastine
124. Vinorelbine
125. Vorinostat
126. WH-4-023
127. XMD8-85
128. Y-39983
129. ZG-10
130. ZM-447439
131. Zibotentan
132. piperlongumine
133. selumetinib

EK 4

1. BRD-A01145011
2. BRD-A02481876
3. BRD-A09722536
4. BRD-A12230535
5. BRD-A13122391
6. BRD-A15100685
7. BRD-A18763547
8. BRD-A22997170
9. BRD-A25004090
10. BRD-A28105619
11. BRD-A28746609
12. BRD-A31107743
13. BRD-A34462049
14. BRD-A34817987
15. BRD-A35108200
16. BRD-A35588707
17. BRD-A36275421
18. BRD-A36318220
19. BRD-A36630025
20. BRD-A38030642
21. BRD-A41692738
22. BRD-A56592690
23. BRD-A59431241
24. BRD-A62182663
25. BRD-A63646118
26. BRD-A67097164
27. BRD-A67788537
28. BRD-A68631409
29. BRD-A70155556
30. BRD-A75817871
31. BRD-A76279427
32. BRD-A80213327
33. BRD-A81541225
34. BRD-A83255679
35. BRD-A93255169
36. BRD-A94377914
37. BRD-K00088062
38. BRD-K00317371
39. BRD-K00627859
40. BRD-K01121114
41. BRD-K01436366
42. BRD-K01567962
43. BRD-K01737880
44. BRD-K01877528
45. BRD-K02113016
46. BRD-K02130563
47. BRD-K02407574
48. BRD-K02492147
49. BRD-K02526760
50. BRD-K02965346
51. BRD-K03406345
52. BRD-K03449891
53. BRD-K04466929
54. BRD-K04623885
55. BRD-K04923131
56. BRD-K05402890
57. BRD-K05649647
58. BRD-K05653692
59. BRD-K06593056
60. BRD-K06750613
61. BRD-K06792661
62. BRD-K06854232
63. BRD-K07303502
64. BRD-K08417745
65. BRD-K08589866
66. BRD-K09485525
67. BRD-K09499853
68. BRD-K09778810
69. BRD-K09907482
70. BRD-K09951645
71. BRD-K10466330
72. BRD-K10705233
73. BRD-K10882151
74. BRD-K11533227
75. BRD-K12040459
76. BRD-K12184916
77. BRD-K12343256
78. BRD-K12502280
79. BRD-K12994359
80. BRD-K13032584
81. BRD-K13044802
82. BRD-K13049116
83. BRD-K13566078
84. BRD-K13999467
85. BRD-K14844214
86. BRD-K15108141
87. BRD-K15318909
88. BRD-K15563106
89. BRD-K15600710
90. BRD-K16189898
91. BRD-K16478699
92. BRD-K17140735
93. BRD-K17349619
94. BRD-K17743125
95. BRD-K17953061
96. BRD-K19295594
97. BRD-K19352500
98. BRD-K19416115
99. BRD-K19477839
100. BRD-K19540840
101. BRD-K19687926
102. BRD-K19796430
103. BRD-K19894101
104. BRD-K20285085
105. BRD-K20755323
106. BRD-K22134346
107. BRD-K22828899
108. BRD-K23984367
109. BRD-K24132293
110. BRD-K24690302
111. BRD-K24844714
112. BRD-K25311561
113. BRD-K25340465
114. BRD-K25737009
115. BRD-K26664453
116. BRD-K26818574
117. BRD-K27224038
118. BRD-K27986637
119. BRD-K28360340
120. BRD-K28428262
121. BRD-K28907958
122. BRD-K29313308
123. BRD-K29458283
124. BRD-K29905972
125. BRD-K30064966
126. BRD-K32330832
127. BRD-K32536677
128. BRD-K33106058
129. BRD-K33379087
130. BRD-K33514849
131. BRD-K33583600
132. BRD-K34157611
133. BRD-K34581968
134. BRD-K35520305
135. BRD-K35708212
136. BRD-K35716340
137. BRD-K35723520
138. BRD-K35960502
139. BRD-K37392901
140. BRD-K37720887
141. BRD-K37798499
142. BRD-K37865504
143. BRD-K38985961
144. BRD-K40853697
145. BRD-K43149758
146. BRD-K43389698
147. BRD-K44241590
148. BRD-K45401373
149. BRD-K45681478
150. BRD-K47150025
151. BRD-K47335880
152. BRD-K47503321
153. BRD-K49075727
154. BRD-K49290616
155. BRD-K49328571
156. BRD-K49456190
157. BRD-K50128260
158. BRD-K50140147
159. BRD-K50168500
160. BRD-K50799972
161. BRD-K51490254
162. BRD-K51575138
163. BRD-K51781482
164. BRD-K52075040
165. BRD-K52313696
166. BRD-K52836380
167. BRD-K52911425
168. BRD-K53414658
169. BRD-K53792571
170. BRD-K53855319
171. BRD-K53903639
172. BRD-K53972329
173. BRD-K54256913
174. BRD-K55116708
175. BRD-K55187425
176. BRD-K55478147
177. BRD-K55591206
178. BRD-K55696337
179. BRD-K56277358
180. BRD-K56301217
181. BRD-K56343971
182. BRD-K57080016
183. BRD-K58550667
184. BRD-K58772419
185. BRD-K59369769
186. BRD-K59456551
187. BRD-K59962020
188. BRD-K60219430
189. BRD-K60230970
190. BRD-K60866521
191. BRD-K61323504
192. BRD-K61662457
193. BRD-K61829047
194. BRD-K62965247
195. BRD-K63068307
196. BRD-K63195589
197. BRD-K63431240
198. BRD-K63923597
199. BRD-K64052750
200. BRD-K64606589
201. BRD-K64610608
202. BRD-K64634304
203. BRD-K64642496
204. BRD-K64785675
205. BRD-K64800655
206. BRD-K64890080
207. BRD-K66175015
208. BRD-K66874953
209. BRD-K67298865
210. BRD-K67566344
211. BRD-K67578145
212. BRD-K67844266
213. BRD-K67868012
214. BRD-K68065987
215. BRD-K68143200
216. BRD-K68548958
217. BRD-K69608737
218. BRD-K69840642
219. BRD-K69932463
220. BRD-K70401845
221. BRD-K71035033
222. BRD-K71726959
223. BRD-K71935468
224. BRD-K72264770
225. BRD-K72420232
226. BRD-K73261812
227. BRD-K73397362
228. BRD-K74065929
229. BRD-K74148702
230. BRD-K74236984
231. BRD-K74514084
232. BRD-K75295174
233. BRD-K75308783
234. BRD-K75430629
235. BRD-K76674262
236. BRD-K76703230
237. BRD-K76964878
238. BRD-K77908580
239. BRD-K78431006
240. BRD-K78599730
241. BRD-K78659596
242. BRD-K79254416
243. BRD-K79877282
244. BRD-K80672993
245. BRD-K81418486
246. BRD-K81473043
247. BRD-K81528515
248. BRD-K81651477
249. BRD-K82109576
250. BRD-K82746043
251. BRD-K83213911
252. BRD-K83289131
253. BRD-K83336168
254. BRD-K84937637
255. BRD-K85133207
256. BRD-K85606544
257. BRD-K86535717
258. BRD-K86574132
259. BRD-K86930074
260. BRD-K87142802
261. BRD-K87737963
262. BRD-K88510285
263. BRD-K88544581
264. BRD-K88742110
265. BRD-K89162000
266. BRD-K89329876
267. BRD-K89391146
268. BRD-K89692698
269. BRD-K89732114
270. BRD-K90370028
271. BRD-K90382497
272. BRD-K92093830
273. BRD-K92241597
274. BRD-K92428232
275. BRD-K92723993
276. BRD-K92991072
277. BRD-K93095519
278. BRD-K93176058
279. BRD-K93367411
280. BRD-K93754473
281. BRD-K93918653
282. BRD-K94991378
283. BRD-K96431673
284. BRD-K96799727
285. BRD-K98404142
286. BRD-K99749624

ÖZGEÇMİŞ

Ad-Soyad	: Ertan Tolan
Uyruğu	: T.C.
Doğum Tarihi ve Yeri	: 18.03.1991 - Aydın
E-posta	: e.tolan@etu.edu.tr

ÖĞRENİM DURUMU:

- **Lisans** : 2014, TOBB ETU, Bilgisayar Mühendisliği

MESLEKİ DENEYİM VE ÖDÜLLER:

Yıl	**Yer**	**Görev**
2014-2016	TOBB ETU	Burslu Yüksek Lisans Öğrencisi

YABANCI DİL: İNGİLİZCE

TEZDEN TÜRETİLEN YAYINLAR, SUNUMLAR VE PATENTLER:

- E. Tolan and M. Tan. Anti-cancer drug activity prediction by ensemble learning. In International Conference on Knowledge Discovery and Information Retrieval, 2016.