TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KİMYASALLARIN GEN DÜZENLEYİCİ ETKİLERİNİN TAHMİNİ İÇİN TRANSFER ÖĞRENİMİ

YÜKSEK LİSANS TEZİ

Bahattin Can MARAL

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Doç. Dr. Mehmet TAN

NİSAN 2022

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.

Bahattin Can MARAL

İMZA

ÖZET

Yüksek Lisans Tezi

KİMYASALLARIN GEN DÜZENLEYİCİ ETKİLERİNİN TAHMİNİ İÇİN TRANSFER ÖĞRENİMİ

Bahattin Can MARAL

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

Tez Danışmanı: Doç. Dr. Mehmet TAN

Tarih: NİSAN 2022

Kemogenomik, ilaç tasarımına ve taramaya yardımcı olmak amacıyla biyolojik hedeflerin kimyasal bileşiklere genomik ve/veya proteomik reaksiyonunun incelenmesidir. Kemogenomikteki birçok zorluktan biri, gerçek yaşam deney verilerine bağımlılıktan kaynaklanmaktadır; farklı kimyasal bileşiklerin ve ilaç hedeflerinin kombinasyonu, gerçekçi olmayan sayıda olası deney yaratır ve bu da belirli kimyasallara ve hedeflere yönelik önyargılı veri kümeleriyle sonuçlanmaktadır. Yapay öğrenmedeki son gelişmeler, bu veri kümelerinin sınırlarını kolayca zorlayan güçlü modellerin aşırı doygunluğuyla sonuçlanmıştır. Bu yatkınlıkların etkilerini nötrlemek için, benzer problemlerden bilgi edinme yöntemi olan transfer öğrenmeyi kullanmaktayız.

Kemogenomik veri setlerindeki en önemli yanlılık, ilaç hedeflerine yönelik olandır. Bazı hücre dizilerinin erişebilirliği ve önemi, bu deneyler için bir ilaç hedefi olarak kullanılma şansını büyük ölçüde artırırken, diğerlerinin yapay öğrenme modellerini eğitmek için ancak yeterli verisi vardır.

Derin Bileşik Profil Oluşturucu (DeepCOP) üzerinde yapılan çalışmayı temel olarak kullanırken, transfer öğreniminin, çeşitli ilaç hedeflerinin eğitilebilirliğini büyük ölçüde artırdığını deneysel olarak göstermekteyiz. Deneyler için kullanılan model yapısı değiştirilmemiştir. DeepCOP'da kullanılan veri bölme yöntemine ek olarak iki yöntem daha eklenmiştir.

Deneylerimiz transfer öğrenmenin basit yöntemlerinden biri olan parametre tabanlı transfer öğrenimine odaklanırken, ROC eğrisi altında kalan alan puanlarında %22,81'e varan ve ortalama %9,00 iyileşme göstermiştir; bununla birlikte hiperparametre optimizasyonu uygulandığı ve transfer kaynağı olarak doğru hücre hattı seçildiğinde bu iyileşmelerin arttırılabileceğine yönelik potansiyel göstermiştir.

Anahtar Kelimeler: Transfer öğrenimi, Kemogenomik, Alan uyarlaması.

ABSTRACT

Master of Science

TRANSFER LEARNING FOR PREDICTING GENE REGULATORY EFFECTS OF CHEMICALS

Bahattin Can MARAL

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

Supervisor: Doç. Dr. Mehmet TAN

Date: April 2022

Chemogenomics is the study of the genomic and/or proteomic reaction of biological targets to chemical compounds, with the goal of aiding drug design and screening. One of the many difficulties in chemogenomics comes from the dependency on real-life experiment data; the combination of different chemical compounds and drug targets creates an unrealistic number of possible experiments, which results in datasets that are biased towards certain chemicals and targets. The recent developments in machine learning resulted in an over-saturation of powerful models that easily pushed the limits of these datasets. To undo the effects of these biases, we employ transfer learning, the method of leveraging knowledge from similar problems.

The most important bias of chemogenomics datasets is the bias towards drug targets. The availability and significance of certain cell lines greatly increase the chance of it being used as a drug target for these experiments, while others have barely enough data to train machine learning models.

We experimentally demonstrate that transfer learning greatly increases the trainability of various drug targets, while using the work done on the Deep Compound Profiler (DeepCOP) as a basis. While focused on one of the simple methods of transfer learning, our experiments showed up to 22.81% and an average of 9.00% improvement on the area under ROC curve scores and showed great potential to be improved upon if accompanied by hyperparameter optimization and correct cell line as the transfer source.

Keywords: Transfer learning, Chemogenomics, Domain adaptation.

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren hocam Doç. Dr. Mehmet Tan'a, kıymetli tecrübelerinden faydalandığım TOBB Ekonomi ve Teknoloji Üniversitesi Bilgisayar Mühendisliği Bölümü öğretim üyelerine, beraber çalışmaktan büyük keyif aldığım Bilgisayarlı Biyoloji ve Makine Öğrenme Laboratuvarı'ndaki çalışma arkadaşlarıma, destekleriyle her zaman yanımda olan annem Canan MARAL, babam Gürsel MARAL, ve kardeşim Bahadır MARAL'a teşekkürlerimi sunarım.

İÇİNDEKİLER

ŞEKİL LİSTESİ

ÇİZELGE LİSTESİ

KISALTMALAR

AUC : Area Under the ROC Curve
DeepCOP : Deep gene COmpound Profiler
ECFP : Extended-Connectivity Fingerprints
FCFP : Functional-Class Fingerprints
LINCS : Library of Integrated Network-based Cellular Signatures
ROC : A Receiver Operator Characteristic
TL : Transfer Learning

1. GİRİŞ

Yapay öğrenme alanında sıkça görülen gelişimlerden dolayı bilimsel araştırma için üretilen veri setlerinin sınırları her geçen gün zorlanmaktadır. Bu gelişimlerin transfer öğrenmede tanımladığımız etki alanlarına direkt olarak etkisi olduğundan, bu alanlar üzerinde eğitilen yeni yapay öğrenim modelleri; çoklu görev model eğitimi, model toplulukları, transfer öğrenimi gibi meta-yöntemleri geride bırakmaktadır. Bu durumun en büyük etkenlerinden biri meta-öğrenim yöntemlerinin kurulumu ve optimize edilmesinin, yeni çıkan modellerle karşılaştırıldığında daha zorlayıcı olması ve her durum için izlenebilecek bir yol haritasının olmamasıdır.

Meta-öğrenim yöntemleri var olan veri setlerinden mümkün olan en fazla bilgiyi çıkartmamızı sağlarken, aynı zamanda geleneksel yapay öğrenme yöntemlerine göre daha komplike ve uyumsallık gerektiren problemlerde daha üstün başarı elde etmekte. Buna örnek olarak süper-çözünürlük problemi örnek olarak verilebilir.

Süper-çözünürlük, yapay öğrenmenin; bilgisayarla görü ve sinyal işleme kategorilerinde, özellik artırma ve netleştirme amacı ile, düşük çözünürlüklü bir görüntü veya sinyal verisinin görüntü kalitesi bozulmadan çözünürlüğünün arttırılması problemine verilen addır. Süper-çözünürlük veri setleri diğer problemlere göre daha hızlı ve ucuza oluşturulabildiği için bu problem yeterince karmaşık bir derin öğrenme metodu için kolay gözükebilir. Fakat son yıllarda düzenlenen konferans ve yarışmalara bakıldığında bu alanın GAN [1] modelleri tarafından domine edildiği görülecektir[2, 3]. Bu üstünlüğün sebebi GAN modellerinin, geleneksel muadillerinin etki alanı sonsuzca geniş süper-çözünürlük problemine uyum sağlayamamasından kaynaklıdır.

Kimyasal bileşikleri ve bunların çeşitli insan hücreleri üzerindeki etkilerini inceleyen biyoinformatiğin son alanı olan kemogenomik[4]; ideal ilaç keşif sürecinin yolunu açmaktadır. Herhangi bir ilaç keşif yönteminin en çok zaman alan kısımlarından biri, yeni ilaç adaylarının üretilmesi ve elenmesi sürecidir, diğer bir adıyla Yüksek Verimli Tarama (YVT, *-ing. HTS*)[5]. Yapay öğrenmenin gelişiminden önce, YVT süreci robotlar ve son derece hassas sensörlerin yardımıyla otomatikleştirilirdi. Aday eleme süreci tamamen fizikseldi ve bu nedenle pahalıydı.

Derin öğrenme gibi güçlü yapay öğrenme teknikleri geldiğinde, kemogenomik simüle edilmiş otomasyon[6, 7] gerektiren en pahalı süreçlerden biriydi. O zamandan beri, kemogenomik, özellikle standartlaştırılmış genomik veri setlerinin kullanıma sunul-

masından beri hızla gelişmektedir [8]. Bu veri kümelerinin çoğu, yalnızca çok çeşitli kimyasal tepkiler içermekle kalmaz, aynı zamanda bunları hücre hatlarına, deney sürelerine ve kimyasal sentezlenebilirliğe göre sınıflandırır. Farklı hücre hatları üzerindeki kimyasal etkileri neredeyse hiç maliyet olmadan simüle etme yeteneği, aday ilaçların potansiyel yan etkilerini erkenden belirleyerek ilaç keşif sürecine de yardımcı olmuştur.

Günümüzde yürütülen kemogenomik çalışmalar Temel Bilgiler2 başlığı altında ayrıntıladığımız ilaç keşif sürecini bilgisayar desteğiyle hızlandırmak ve süreç maliyetini düşürmek amacıyla yapılmaktadır. Laboratuvar ortamında yapılan deneyleri bilgisayar üzerinde simüle edebilmek için gereken şartlar:

- Kimyasal bileşiklerin temsili
- Hücre dizilerinin temsili
- Hücre dizilerinin kimyasallara olan tepkilerinin temsili
- Yeni kimyasalların oluşturabileceği etkilerin tahmini
- İstenen etkileri oluşturabilecek sentezlenebilir kimyasal tahmini

olarak sıralanabilir.

Kemogenomik çalışmaların ilaç keşif sürecinin yerini alması içinse bu şartların herbirinin mükemmele yakın şekilde sağlanması gerekmekte. Yapılan çalışmalar bu şartların bir veya birkaçını hedef alabiliyor.

Kemogenomiğe yönelik en son derin öğrenme yaklaşımlarından biri, Derin Gen Bileşik Profil oluşturucu (DGBPo, *-ing. DeepCOP*) adlı çalışmadır. DeepCOP, Entegre Ağ Tabanlı Hücresel İmzalar Kitaplığı (EATHİK, *-ing. LINCS*) L1000 [9] transkriptomik veri kümesi üzerinde eğitilmiş bir derin öğrenme modelidir. Araştırmada, araştırmacılar seçilen her hücre dizisi için çok katmanlı algılayıcı (ÇKA, *-ing. MLP*)[10] biçiminde iki derin öğrenme modeli eğitmektedir. Bu modellerden biri, genlerin yukarı regülasyonunu, diğeri ise aşağı regülasyonunu ikili sınıflandırma biçiminde tahmin etmektedir. Çalışma, özellik üretimine odaklanmasıyla ve basit derin öğrenme metoduyla kendini farklılaştırmaktadır. Temelinde, hücre dizilerinin temsilini iyileştirerek kimyasal tepkilerini eniyilemeye çalışan bir araştırmadır.

Kimyasal tepkimelerin standartlaştırılıp kategorize edildiği L1000 benzeri veri setleri laboratuvar ortamında yapılan deneyler ile oluşturulmaktadır. Bu nedenle veri setlerindeki deneyler her hücre dizisi ve kimyasal için çoğu zaman eşit bir şekilde dağılmamaktadır. Bu durum bazı hücre dizileri için bulunan verilerin kimyasal etki tahmini yapmak için

yetersiz miktarda olması veya eğitilen modellerin genelleştirilememesi ile sonuçlanmaktadır.

Yapay öğrenmede transfer öğrenimi, bir problemi farklı fakat ilgili bir problem üzerinde çözerken daha önce kazanılan bilgilerin kullanılması olarak tanımlanır. Bu çalışma için, farklı problemler farklı hücre hatlarına karşılık gelmektedir. Diğer hücre dizilerinden elde edilen verilerin kullanılması, hücre hatlarına özgü kimyasal bileşiklerle deneyler yapılması nedeniyle daha genel bir modelin eğitilmesi avantajını da beraberinde getirir.

Bu çalışmada ise, bu eğilimlerin etkilerini nötrlemek, yeterli sayıda verisi olmayan hücre dizileri ile kullanılabilir modeller eğitebilmek için transfer öğreniminden yararlanmaktayız.

Bu çalışma ile literatüre sunduğumuz katkılar aşağıda listelenmektedir:

- DeepCOP çalışmasına gen pertürbasyonu üzerinde eğitilen model performansını daha iyi temsil eden 2 veri bölüm yöntemi ekleyerek çalısmanın geliştirdiği gen temsili yönteminin gerçekçi sonuçları alınmıştır.
- DeepCOP'ta kullanılan 6 hücre dizisindeki model eğitimine ek olarak, çalışmada veri azlığı nedeniyle elenmiş 17 hücre dizi verisi üzerinde de eğiterek DeepCOP'un bu hücre dizilerindeki performansı ölçülmüştür.
- Kenogenomik yapay öğrenme deneylerinde rastgele veri bölümünün ilaç bazında veri bölümü ile karşılaştırmasını yapıp, avantaj ve dezavantajları listelenmiştir.
- Rastgele, soğuk, ve transfer ilaç ayrımı adı altında 3 farklı veri bölümü üstünde transfer öğrenimin kemogenomik uygulamalar üzerindeki etkisini ve kullanılabilirliğini deney sonuçları ile gösteriyoruz. Bu deneyler sonucunda ayrımlar üzerinde sırasıyla ortalama %4,52, %9,00 ve %0,69 AUC skoru iyileşmeleri gözlemlenmiştir.
- Transfer ilaç ayrımı isimli yeni bir veri bölüm yöntemi tanıtarak, bu ayrımın soğuk ilaç ayrımı ile karşılaştırıldığında farklarını ve bu karşılaşımdan çıkarılabilecek sonuçları örneklenmiştir.
- 6 farklı hücre dizisi üzerinde eğitilen modellerden elde edilen bilgiler kullanılarak, bu bilgilerin 23 hücre dizini üzerinde eğitilen modeller üzerindeki etkilerini gözlemlenmiştir.

2. TEMEL BİLGİLER

2.1 Yapay Sinir Ağları ve Çok Katmanlı Algılayıcılar

Çok katmanlı bir ileri beslemeli yapay sinir ağı (YSA, *-ing. ANN*) [11], çok katmanlı algılayıcı (ÇKA, *-ing. MLP*) [10] olarak bilinmektedir. MLP adı belirsizdir; herhangi bir ileri beslemeli YSA'ya atıfta bulunmak için kullanılabilmektedir veya birçok algılayıcı katmanından oluşan (eşik aktivasyonlu) ağlara atıfta bulunabilmektedir. Çok katmanlı algılayıcılar, özellikle tek bir gizli katmana sahip olanlar, yaygın olarak "vanilya" sinir ağları olarak adlandırılmaktadır. Bir ÇKA'da en az üç seviye düğüm vardır: bir giriş katmanı, bir gizli katman ve bir çıkış katmanıdır. Giriş düğümleri dışındaki her düğüm, doğrusal olmayan aktivasyon fonksiyonuna sahip bir nörondur. Geri yayılım, eğitim sırasında MLP tarafından kullanılan denetimli bir öğrenme tekniğidir. MLP, çok sayıda katmanı ve doğrusal olmayan aktivasyonu ile doğrusal bir algılayıcıdan ayırt edilmektedir. Doğrusal olarak ayrılamayan veriler arasındaki farkı ayırt edebilmektedir.

2.2 Transfer Öğrenimi

(Aktarım yoluyla öğrenme), bir görevde kullanılan bir modelden elde edilen bilgilerin başka bir faaliyet için temel olarak kullanıldığı bir yapay öğrenme stratejisidir. Yapay öğrenme algoritmaları, geçmiş verileri girdi olarak kullanarak tahminlerde bulunmaktadır ve yeni çıktı değerleri üretmektedir. Genellikle bir seferde tek bir iş yapmak için eğitilmektedirler.

Kaynak görev, eğitebilirliği yüksek, elde edilen bilginin hedef göreve aktarılacağı etki alanı olarak tanımlanır. Hedef görev, bir kaynak görevden bilgi aktarımının bir sonucu olarak daha iyi öğrenmenin gerçekleştiği bir görevdir. Transfer öğrenimi sırasında, kaynak görevde kazanılan bilgi ve kaydedilen hızlı ilerleme, yeni bir hedef görevin öğrenilmesine ve geliştirilmesine katkı sağlamaktadır.

Aktarılan bilginin hedef görev performansında düşüşe neden olmasına Negatif Transfer adı verilmektedir. Transfer öğrenme yöntemlerini kullanırken, en zor konulardan biri, ilgili görevler arasında pozitif transfer sağlamak ve ilişkisiz etkinlikler arasında negatif transferden kaçınmaktır.

Yapay öğrenmede sıkça karşılaşılan bir durum ile örneklendirirsek: Optik Karakter

Tanıma (OKT, *-ing. OCR*) yapay öğrenmenin alt dalı olan bilgisayarla görü biliminin ilk çözmeye çalıştığı problemlerden biridir. OCR, taranmış bir belgeden, bir belgenin fotoğrafından, bir sahne fotoğrafından veya bir görüntünün üzerine bindirilmiş altyazı metninden, daktilo edilmiş, elle yazılmış veya basılı metin görüntülerinin yapay tarafından kodlanmış metne elektronik veya mekanik olarak dönüştürülmesi olarak tanımlanmaktadır.

OCR problemi üzerine çoğu dil için %100 başarı ile çalışabilen modeller eğitilmiş olup, dillerin büyük çoğunluğu için çözülmüş bir problem olarak anılmaktadır. İngilizce makale çıktılarından elde edilmiş çok sayıda etiketli veri ile eğitilmiş bir OCR yapay öğrenme modeli ele alalım: Eğitim verileri ve hedef verilerin her ikisi de İngilizce metinlerden türetilmişse, iyi tahmin sonuçları elde etmek için geleneksel yapay öğrenme teknikleri kullanılmaktadır. Ancak, eğitim verilerinin İngilizce metinlerden ve hedef verilerin Bambara metinlerinden oluşması durumunda, diller arasındaki farklılıklar nedeniyle tahmin sonuçlarının düşmesi muhtemeldir. İngilizce ve Bambara dili gramer ve telaffuz açısından benzemeseler de veri setleri bir takım ortak özelliklere sahiptir: Her iki dil de latin alfabesinden türemiş alfabelere sahiptir ve her iki veri de benzer şekillerde üretilmiştir. Bu iki alan birbiriyle ilişkili olduğundan, transfer öğrenimi, hedef öğrenicinin sonuçlarını potansiyel olarak iyileştirmek için kullanılabilir. Bir transfer öğrenme ortamındaki veri alanlarını görüntülemenin alternatif yolu, eğitim verilerinin ve hedef verilerin, yüksek seviyeli bir ortak alan ile bağlantılı farklı alt alanlarda bulunmasıdır. Örneğimiz için problemlerimiz optik karakter tanıma probleminin alt alanlarıdır. Üst düzey ortak alan, alt alanların nasıl ilişkili olduğunu belirlemektedir.

2.2.1 Transfer öğrenme kategorileri

Transfer öğrenme yöntemleri birden fazla kritere göre kategorize edilebilir.

2.2.1.1 Probleme göre sınıflandırma

Örneğin, transfer öğrenme problemleri üç tipte sınıflandırılabilir: transdüktif, tümevarım ve denetimsiz transfer öğrenme. **Transdüktif transfer öğrenimi**, geniş anlamda, etiket bilgisinin yalnızca kaynak etki alanından geldiği durumları ifade etmektedir. Hedef etki alanı örnekleri için etiket bilgisi mevcutsa, senaryo **tümevarımsal aktarım öğrenimi** olarak sınıflandırılabilir.

Hem kaynak hem de hedef etki alanları için etiket bilgisi bilinmediğinde, duruma **denetimsiz transfer öğrenimi** adı verilir. Transdüktif ve tümevarım transfer öğrenme

regresyon ve kategorizasyon alanlarında kullanılırken; denetimsiz transfer öğrenmeden kümeleme ve boyutsal küçülme problemlerinde faydalanılmaktadır.

2.2.1.2 Çözüm yöntemine göre sınıflandırma

Transfer öğrenme yöntemleri ise dört gruba ayrılabilir: örnek tabanlı, özellik tabanlı, parametre tabanlı ve ilişkisel tabanlı yaklaşımlar.

Örnek tabanlı yaklaşımlar, temelde girdi ağırlığına dayanan stratejilerden oluşmaktadır. Çok sayıda etiketlenmiş kaynak etki alanı ve az sayıda hedef etki alanı örneğinin bulunduğu ve etki alanlarının yalnızca marjinal dağılımlarda farklılık gösterdiği basit bir senaryo düşünülürse: Örneğin piyasaya yeni sürülecek bir cep telefonu üretim hattı üzerinde bilgisayarla görü kullanılarak hatalı üretim tespiti yapabilen bir yapay öğrenme modeli eğitmeye çalıştığımızı düşünelim. Elimizdeki verinin sınırlı olduğu ve veri üretiminin maliyetli olduğunu, problemimize en yakın veri setinin bir önceki telefon modelinden oluştuğunu varsayalım. Elimizdeki bütün veriyi direkt olarak yeni etki alanında kullanmak marjinal farklılıklardan dolayı başarısız olabilir. Böyle bir problem için basit bir çözüm azınlıkta olan etki alanına dahil olan örneklerin kayıp fonksiyonu üzerindeki ağırlığını arttırmak olabilir, bu çözüm yöntemi örnek tabanlı kategorisi altında bulunan örnek ağırlandırmaya bir örnektir.

Örnek tabanlı transfer öğrenme kategorisi altındaki bir diğer yöntem ise etki alanı ağırlandırmasıdır. Bu yöntem etki alanlarına farklı ağırlıklar yükleyerek etki alanları arasındaki farkların modele etkisini dengelemektedir. Bu çözüm yöntemi ağırlıkla birbirine çok daha yakın birçok etki alanı arasında çoklu görev öğrenme modelleri için kullanılır.

Özellik tabanlı yaklaşımlar genellikle özellik transformasyonu metodlarından oluşmaktadır. Birden fazla etki alanında mevcut bir zaman serisi problemi senaryosu ele alalım. Aynı iklim özelliklerine sahip farklı enlemlerde bulunan iki şehir üzerinde iklim tahmini yaptığımızı varsayalım. Benzer iklime sahip olsalar bile farklı enlemlerde bulunmalarından dolayı sıcaklık değişimleri benzerlik gösterirken ortalama sıcaklıkları farklı olacaktır. Bu durumda kaynak şehirden elde edilen veriler hedef şehrin verilerine göre normalize edilerek, büyük bir hassasiyet kaybı yaşanmadan hedef şehir için eğitilecek modelin başarısı arttırılabilir.

Bir diğer özellik tabanlı yaklaşım ise özellik artırımıdır. Bu yöntemde azınlıkta olan hedef etki alanına ait örnekler çeşitli veri türetme yöntemleri ile çoğaltılabilir, tekrarlanabilir, veya istiflenebilir.

Kümeleme yöntemleri de özelllik tabanlı yaklaşımlara örnek verilebilir. Özellik tabanlı yaklaşımlar altındaki kümeleme yöntemleri; örnek kümeleme, özellik kümeleme, ve çıktı kümeleme olarak sıralanabilir. Kümeleme metodları kaynak ve hedef etki alanına ait bilgileri toplam kaybı minimize edecek şekilde bir araya getirmek için kullanılmaktadırlar.

Özellik tabanlı yaklaşımların son alt kümesi ise özellik eşleme üzerine geliştirilmiş çözüm yöntemleridir. Özellik eşleme yöntemleri örnek özelliklerini Temel Bileşenler Analizi (TBA, *-ing. PCA*), Bağımsız Bileşen Analizi (BBA, *-ing. ICA*), Faktör Analizi (FA), Otomatik Kodlayıcılar, vb. yöntemler yardımıyla daha küçük bir boyuta eşlenir. Bu eşleşme sayesinde özellikler özetlenerek etki alanları arasındaki farklılıklar azaltılmaktadır. Eşleşme sonucunda elde edilen özet veri üzerinden model eğitilebilir veya veri geri dönüştürülerek orijinal boyutu üzerinden eğitilebilir.

Parametre tabanlı yaklaşımlar, bilgiyi model/parametre düzeyinde aktarmaktadır. Benzer bir kaynk etki alanı üzerinde eğitilmiş modelin parametreleri hedef etki alanında baz alınmaktadır. **İlişkisel tabanlı transfer öğrenme yaklaşımları** temel olarak ilişkisel alanlardaki problemlere odaklanmaktadır. Bu tür yaklaşımlar, kaynak alanda öğrenilen mantıksal ilişkiyi veya kuralları hedef alana aktarmaktadır.

Parametre ve ilişkisel tabanlı transfer öğrenme yöntemlerinin kullanıldığı çalışmalarda diğer transfer öğrenme yöntemlerinin daha iyi başarım arttırdığı gözlemlendiğinden sadece parametre veya ilişkisel transfer öğrenme yöntemi kullanan çalışmalara ender rastlanmaktadır.

2.2.1.3 Modele göre sınıflandırma

Transfer öğrenimi model tabanlı sınıflandırıldığında iki ana kategoriye ayrılır: strateji ve hedef tabanlı transfer öğrenimi yöntemleri.

Strateji tabanlı yöntemler kendi içerisinde birçok alt kategoriye ayrılabilir:

- **Model Kontrolü** farklı etki alanları üzerinde eğitilen modellerin birbirlerinin sonuçlarını etkilemesidir.
- **Parametre Kontrolü** farklı etki alanları üzerinde eğitilen modellerin arasında parametre paylaşımı veya sınırlaması yapmasıdır.
- **Model Toplulukları** birçok modelin sonucu değerlendirilerek ortak bir tahmine ulaşılma yöntemlerine verilen genel addır.

- **Derin Öğrenme Yöntemleri** Farklı etki alanlarına uyum sağlayabilen otomatik kodlayıcılar [12], üretken düşmanlık ağları (ÜDA, *-ing. GAN*) [13] gibi modellerin kullanıldığı alt kümedir.

2.3 Kemogenomik

Kemogenomik, biyolojik bir sistemin kimyasal maddelere karşı genomik ve/veya proteomik tepkisinin veya izole moleküler hedeflerin bunlarla etkileşime girme yeteneğinin incelenmesidir. Kemogenomik, biyolojik hedefler ve fizyolojik olarak aktif kimyasallar için aynı anda geniş kimyasal bileşik koleksiyonlarını taramayı içermektedir. Tahmine dayalı kemogenomik teknikler, yüzlerce ilaç yanıt profilini analiz ederek gen-bileşik yanıt ilişkilerini tanımlamayı ve ikincil hedef olarak yeni terapötik bileşikler keşfetmeyi amaçlamaktadır.

2.3.1 İlaç keşif süreci

Yeni bir ilacın keşfi ve geliştirilmesi sırasında, sürecin çeşitli aşamalarında çalışılan maddelere çeşitli terimler atanmaktadır. Yalnızca bir madde bir seviyenin gerekliliklerini karşılıyorsa, bir sonraki seviye için teste tabi tutulmaktadır. Bu prosedür, iyi tanımlanmış bir kimyasal yapıya ve saflığa sahip bir kimyasal madde olan bir bileşik ile başlamaktadır. Spesifik bağlanma özellikleri sergileyen ve aktif olan bileşiklere isabet adı verilir ve ayrıca kimyasal kimliklerinin ve saflıklarının doğrulanmasının yanı sıra bir veya daha fazla ikincil tahlilde bağlanma özelliklerinin belirlenmesini içermesi doğrulanmaktadır.

Aday bileşikler üzerindeki takip araştırmaları daha sonra, bir bileşiğin fizikokimyasal özellikleri, ilaç benzerliği ve sentez fizibilitesi gibi pratik faktörleri göz önünde bulunduran karmaşık kemoinformatik stratejileri kullanılarak önceliklendirilmektedir. Bu işlem kurşun optimizasyonu olarak bilinir ve Niceliksel Yapı-Aktivite İlişkisi (QSAR) verilerinin yanı sıra rafine özelliklere sahip benzer sentetik bileşiklerin sentezi kullanılmaktadır. Bu rafine özellikler, biyolojik model sistemlerinde test için uygun ilaç benzeri özelliklerin yanı sıra artırılmış güç ve hedef seçiciliği içermektedir. Gelişmiş özelliklere sahip sentetik analoglar, biyolojik etkilerini ilk kurşun bileşiklerininkilerle karşılaştırmak için tahlile dayalı yüksek verimli taramaya (YVT, *-ing. HTS*) tabi tutulmaktadır. Aşamalı kurşun optimizasyonu, tekrarlanan eşzamanlı HTS turlarından ve kimyasal analog sentezinden kaynaklanmaktadır.

Hücresel veya biyolojik model sistemleri daha sonra keşfedilen isabetleri in vitro ve/veya in vivo doğrulamak ve ayrıca bir bileşiğin etkisinin potansiyelini ve özgüllüğünü

değerlendirmek için kullanılmaktadır.

Klinik çalışmaların 1-3. aşamaları sırasında, seçilen klinik adaylar, belirli terapötik endikasyonlar için toksisite ve etkinlik profillerini değerlendiren önceden tanımlanmış protokollere göre insanlarda test edilmektedir. İlaç keşif ve geliştirme sürecinin amacı, tıbbi bir durumu tedavi etmek için doza bağlı bir şekilde kullanılan, ruhsatlı ve onaylanmış bir kimyasal madde olan bir 'ilaç' üretmektir.

2.3.2 Hücre dizileri

Bir hücre dizisi (veya hücre hattı), yeterli taze ortam ve boşluk verilirse süresiz olarak çoğalmaya devam edecek kalıcı olarak oluşturulmuş bir hücre kültürüdür. Diziler, ölümsüzleştirilebilmeleri bakımından hücre türlerinden farklıdır.

Hücre kültürü ve hücre hatları, spesifik hücrelerde fizyolojik, patofizyolojik ve farklılaşma süreçlerinin incelenmesinde çok önemli hale gelmiştir. Kontrollü ortamlarda hücrenin yapısındaki, biyolojisindeki ve genetik yapısındaki kademeli değişikliklerin çalışılmasını sağlamaktadır. Bu, özellikle farklı hücre tiplerinden oluşan ve tek tek hücrelerin laboratuvar deneyleri ile incelenmesinin imkansız değilse de zor olduğu pankreas gibi karmaşık dokular için yararlıdır. Doğal özelliklerini korurken, tek tek epitel hücrelerini karmaşık dokulardan izole etme ve saflaştırmadaki aşırı zorluk, fizyolojik, biyolojik, büyüme ve farklılaşma özelliklerini anlamamızı engellemiştir. Bu çalışmada kullanılan hücre dizileri Çizelge 2.1'de görülebilir.

2.3.3 Bağlantı haritası

Yan etki tahmin problemine bir çözüm olarak inşa edilen Bağlantı Haritası (BH, *-ing. CMap*) [14], genetik ve farmakolojik bozulmalar ile sistematik bozulmayı temsil eden bir hücresel imzalar kataloğudur. Bu şekilde, yüksek benzerliğe sahip imzalar, faydalı ve önceden tanınmayan bağlantıları temsil edebilir (Örneğin, aynı yolda çalışan iki protein arasında, küçük bir molekül ile protein hedefi arasında veya benzer işleve sahip ancak yapısal farklılık gösteren iki küçük molekül arasında).

2.3.4 LINCS L1000

L1000 [9], CMap'i 1000 faktörü ile ölçekleyen, yüksek verimli, azaltılmış bir temsil ifadesi profili oluşturma yaklaşımıdır.

Çizelge 2.1: Bu çalışmada yer alan hücre hattı kimliklerinin ayrıntıları.

Hücre Hattı	**Birincil Alan**	**Örnek Tipi**	**Alt Tip**
A375	deri	tümör	kötü huylu melanom
A549	akciğer	tümör	küçük olmayan hücre akciğer kanser karsinom
BT20	meme	tümör	karsinom
HA1E	böbrek	normal	normal böbrek
HCC515	akciğer	normal	karsinom
HEK293T	böbrek	normal	embriyonal böbrek
HEPG2	karaciğer	tümör	hepatoselüler karsinom
HL60	hematopoietik, lenfoid doku	tümör	akut miyeloid lösemi, promiyelositik
HS578T	meme	tümör	karsinom
HT29	kalın bağırsak	tümör	kolorektal adenokarsinom
HUH7	karaciğer	tümör	hepatoselüler karsinom
JURKAT	hematopoietik, lenfoid doku	tümör	akut lenfoblastik lösemi, T-hücre
MCF10A	meme	normal	epitel
MCF7	meme	tümör	adenokarsinom
NKDBA	böbrek	normal	böbrek epitel
NOMO1	hematopoietik, lenfoid doku	tümör	akut miyeloid lösemi
PC3	prostat	tümör	adenokarsinom
SKBR3	meme	tümör	adenokarsinom
THP1	hematopoietik, lenfoid doku	tümör	akut miyeloid lösemi, monocytic
U266	kan	tümör	miyeloman, hematopoietik, lymphoid
U937	hematopoietik, lenfoid doku	tümör	lenfoma, B-hücre, non-Hodgkin, histiocytic
VCAP	prostat	tümör	karsinom

L1000, son derece tekrarlanabilir, RNA dizilimi ile karşılaştırılabilir ve ölçülmemiş transkript ekspresyon seviyelerinin yüzde 81'inin hesaplamalı çıkarımı için uygundur.

Genişletilmiş CMap, küçük kimyasal etki mekanizmalarını bulmak, hastalık gen varyasyonlarını işlevsel olarak açıklamak ve klinik deneylere rehberlik etmek için kullanılabilir.

2.3.5 Morgan parmak izleri

İyi bilinen Genişletilmiş bağlantı parmak izlerine (ECFP'ler) veya İşlevsel sınıf parmak izlerine (FCFP'ler) benzer şekilde, Morgan parmak izleri, kimyasal yapıları ikili diziler olarak temsil etmek için kullanılmaktadır.

[15]'in ilk tanıtıldığı andan itibaren, [16]'deki modernleştirilmiş uygulamaya kadar, Morgan parmak izleri bugüne kadarki en yaygın parmak izi alma yöntemlerinden biri olmuştur.

2.3.6 Gen ontoloji terimleri

Gen Ontoloji (GO) [17], yüksek verimli biyolojik veri kümesi analizi ve yorumu için evrensel bir kaynaktır. GO Konsorsiyumu, dünya çapındaki bilimsel kurumlarda bulunan bir dizi farklı grup tarafından yapılan GO'nun geliştirilmesini ve küratörlüğünü denetlemektedir. GO açıklamaları, gen ürünlerini GO terimleriyle ilişkilendirerek biyolojik işlevsel bilgiyi yakalamaktadır.

GO terimlerinin ve gen ürün kayıtlarının hepsinde bilgisayar tarafından okunabilen erişim numaraları olduğundan, bunlar insan tarafından okunabilir etiketleri tutarken çok büyük veri kümelerini analiz etmek için kullanılabilmektedir.

3. GEÇMİŞ ÇALIŞMALAR

Kemogenomik çalışmalar giriş1 başlığı altında maddelendirdiğimiz ilaç keşfi sürecinin farklı adımlarını hedef almaktadır. Bu çalışmaların sonucunda sürecin herhangi bir aşamasının iyileştirilmesi diğer aşamalarda alınan sonuçların potansiyelini direkt olarak etkilemektedir. Bu durumun yarattığı dalgalanma sonucu ortaya çıkan çalışmalar keşif aşamasının temel unsurları üzerine yoğunlaşmaktadır.

Bu sebeple bu çalışmada hedef gösterdiğimiz kimyasalların genler üzerine etkisi aşamasında yapılan çalışmalara [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29] gibi birçok örnek verilebilir.

[18, 30, 21] çalışmaları DeepCOP'a benzer bir şekilde L1000 veri seti kullanılarak yapılan çalışmalarken, [20, 23, 26] DrugBank [31], [22, 24] ise ChEMBL [32] tarafından sağlanan veri tabanı üzerinde; [25] SIDER [33], [27] KIBA [34] veri setlerinde; [28, 29] çalışmaları ise karma veri setleri üzerinde test edilmiştir.

Bu çalışmalar ilaç-gen etkileşimi [18, 30, 21], ilaç-protein etkileşimi [29, 27], yan etki tahmini [25], ilaç-hedef bağlanma afinitesi tahmini [28] gibi alt kategorilere ayrılabilir. Bu çalışmaların ortak noktaları ise kimyasal etki tahminini iyileştirmeye yoğunlaştıkları için bu çalışmalar veri setlerinde deney sayısı yüksek hücre dizileri üzerinde gerçekleştirilmektedir.

Bu sebeple bu alanda yapılan çalışmalarda çoklu görev, transfer öğrenimi gibi meta iyileştirme çalışmaları çok daha nadir görülmektedir. Bu çalışmalara [35, 36, 37, 38, 39] örnek olarak verilebilir.

[39]'da araştırmacılar ChEMBL veri seti üzerinde doğal bileşikler çıkarılmış veri ile eğittikleri MLP modelini kaynak model olarak kullanıp, doğal bileşikler etki alanına transfer öğrenme gerçekleştirmiştir. Bu sayede genelleştirme özelliğini kaybetmemiş ve doğal bileşiklerin gen regülasyonunu tahmin edebilen bir model eğitmiştir.

Model tabanlı transfer öğrenimi örneği verecek olursak: [38]'nın araştırmacıları hem yapay sinir ağı nöron ağırlıkları hem de model topluluğu prensipleri üzerinden kanser ilaçları için transfer öğrenim modelleri eğitmiştir. Kullandıkları model toplulukları içinde bir geleneksel yapay sinir ağı, bir adet LightGBM [40] modeli, ve iki alt-ağa sahip olan bir yapay sinir ağı bulunmaktadır.

Benzer şekilde, [37] no'lu çalışmada araştırmacılar Kanser Genom Projesinin Kanserde İlaç Duyarlılığı (KİD, *-ing. GDSC*) [41] ve Kanser Hücre Hattı Ansiklopedisi (KHHA, *-ing. CCLE*)) [42] veri setlerini özellik-tabanlı iki, model topluluğu ve özellik tabanı bir yöntem olmak üzere toplam 3 farklı yöntemle bir arada kullanmışlardır.

Kanser verisiyle üzerinde çalışan bir diğer çalışmada [35], araştırmacılar verisi az bulunan Çoklu Miyelom Hastalığı için diğer kanser veri setlerinden örnek-tabanlı transfer öğrenme uygulayarak SMOTE [43] yardımıyla Çoklu Miyelom için yeni veri türetmişlerdir. Aynı araştırma ekibi bir sene sonra [36] çalışmalarını hedef etki alanı diğer verisi az kanser tiplerini de kapsayacak şekilde genişletmiştir. Bu çalışmalarda aynı zamanda SVR [44], doğrusal SVM [45], ridge regresyon [46], ve lojistik regresyon [47] gibi model sonuçları beraber kullanılmıştır.

GO terimleriyle çalışılan bir çalışmaya örnek verecek olursak GO-TLM [48] adlı çalışma GO terimleri ile protein hücre altı lokalizasyonunun tahmini yaparken, bilinen protein benzerlikleri arasında model tabanlı transfer öğrenimi uygulayarak; modellerinin başarısını arttırmış ve verilerindeki gürültü ve uç örnekleri elemişlerdir.

Bioenformatik alanında yapılan transfer öğrenimi yapılan çalışmalar, yalnızca insana ait etki alanları arasında bilgi aktarımı yapmakla sınırlı kalmamaktadır. [49] ve [50] çalışmalarında araştırmacılar farelerde yapılan deneyleri kullanarak insan hücreleri üzerinde eğitilen yapay öğrenme modellerinin başarısını arttırmıştır. [50] bununla kalmayıp fareler üzerinde yapılan deneylerin insanlar üzerindeki etkisini tahmin edebilmeyi hedeflemektedir. Bu amacın sebebi bu canlılardan elde edilen deney verisinin ve hücre dizisi miktarının çokluğudur.

4. YÖNTEMLER

4.1 DeepCOP

Önceki bölümlerde kısaca bahsedildiği üzere bu çalışmanın temelinde DeepCOP makalesinde geliştirilmiş çalışma yer almaktadır. DeepCOP'un hedefi, her gen çıktısının daha iyi temsil etmektir. Bunun için çalışmada araştırmacılar biyoenformatik alanında sıklıkla kullanılan Morgan parmak izlerini kimyasalları temsil etmek için, gen ontolojisi konsorsiyumu çalışmasından çıkarttıkları vektörleri ise L1000 veri setindeki hücre dizilerine ait 978 dönüm noktası geninin temsili için kullanmıştır. DeepCOP yapay sinir ağı eğitmek için kullandığı veri setini oluşturmaya LINCS L1000 CMap veri seti ile başlamıştır:

- Öncelikle L1000 verisi standartlaştırılarak z-skorları elde edilmiştir.
- Daha sonra veri setinden 24 saatten kısa süren deneyler çıkartılmıştır, çıkartılan veri toplam verinin yaklaşık %47'sini oluşturmaktadır.
- Kalan veriden ilaç dosajları μM ile ölçülen deneyler (yaklaşık %42'si) seçilmiştir.
- Bu veriden ilaç konsantresi en yüksek olan deneyler seçilmiştir (10μM).
- Elde edilen veri setinden denenmiş kimyasal sayısı 5000'in üzerinde olan 6 hücre dizisi seçilmiştir. Bunlar VCAP, A549, A375, PC3, MCF7, ve HT29 kod adlı hücre dizileridir (bkz. 2.1).

Daha sonra filtrelenmiş L1000 verisinden Morgan ve GO özellik vektörleri elde edilmiştir. Çalışmada, kimyasal yapıyı gen pertürbasyonuyla ilişkilendirmek için RDKIT Açık kaynaklı kemoinformatik araç kiti [51] kullanılarak Morgan tanımlayıcıları üretilmiştir. Geleneksel SMILES formundaki tanımlayıcılar, 19,811 bileşik için 2048 öznitelikli bir bir-elemanı-bir vektör oluşturmak üzere 2 yarıçapla hesaplanmıştır.

OntologyX R paketi kullanılarak, en az 3 dönüm noktası gen ile ilişkili GO terimleri çıkarılmıştır. Daha sonra her bir gen, uygun GO terimleri kullanılarak tanımlanmıştır. Bu GO tanımlayıcılarının bir-elemanı-bir kodlaması, Morgan parmak izlerinin yanında özellik olarak kullanılabilecek 1107 boyutundaki ikili vektörle sonuçlanmıştır.

Elde edilen veriseti ile herbiri 3155 nörondan oluşan 2 gizli katmanlı bir yapay sinir ağı 10-katlamalı rastgele bölünmüş çapraz doğrulama ile eğitilmiştir. Bu deneyler sonucunda Çizelge 5.1 ve Çizelge 5.2'da bulunan TL Yok sütunundaki VCAP, A549, A375, PC3, MCF7, ve HT29 ait sonuçlar elde edilmiştir.

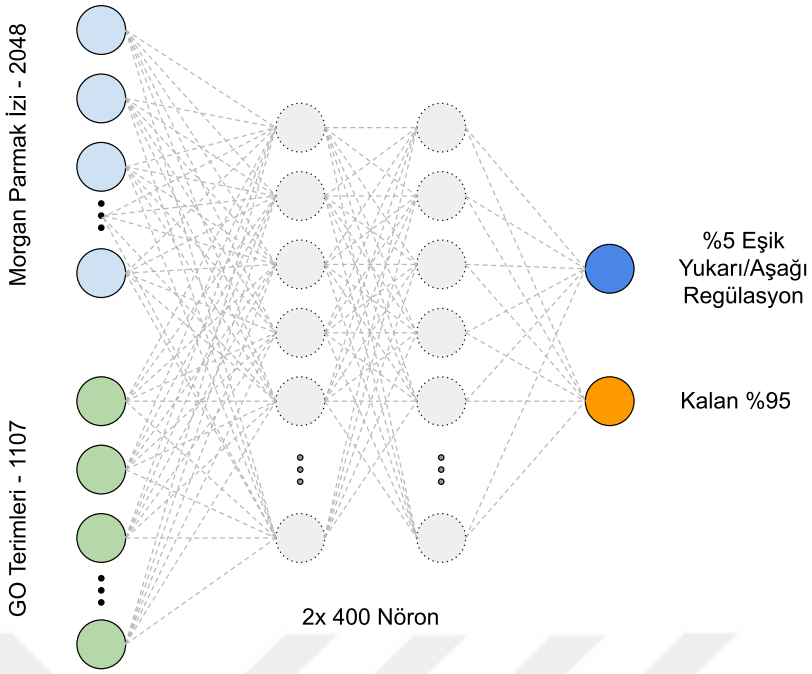
4.2 Önerilen Metot

4.2.1 Veri bölümü

DeepCOP'ta araştırmacılar, her hücre hattı için 10 kesit çapraz doğrulama gerçekleştirmek için bahsedilen veri setini rastgele seçilmiş 10 kesite bölmüştür. Bu yaklaşım, her bir bileşik-gen kombinasyonunun kesit başına benzersiz olmasını sağlamaktadır. Fakat, her ilacın, dönüm noktası genlerinin sayısı ile tekrarlandığı gerçeğini göz ardı etmektedir. Aynı şekilde, her gen de bileşik sayısı kadar tekrarlanmaktadır. Bu nedenle, rastgele bölünmüş 9 kesit üzerinde eğitilmiş bir model, geri kalan kesit için ayrı ayrı girdi öğeleriyle (bileşik ve gen) zaten karşılaşmıştır. Bu, tartışma bölümünde daha ayrıntılı olarak ele aldığımız nedenlerden dolayı, aşırı takmaya neden olabilecek komplikasyonlara yol açmaktadır. Bu çalışmanın eksiksiz olması adına, bu bölme şekli "rastgele-ilaç-bölme" olarak adlandırılarak bu formatta da deneylerimiz tekrarlanmıştır.

Soğuk-İlaç bölme yöntemi, bileşik verileri 10 kesite bölmektedir, ardından her kesiti gen verileriyle doldurmaktadır. Bu, test bölümü her kesitteki ilaçların eğitimli modeller için daha önce karşılaşılmamış olmasını sağlamaktadır. Bu bölme yöntemine "soğuk-ilaç-bölme" adını vermekteyiz. Transfer öğrenimi için, kaynak model %95-%5 rastgele bölmede eğitilmektedir. Sonrasında hedef hücre hattı verileri soğuk-ilaç-bölme kesitlerinde eğitmekteyiz. Soğuk-İlaç bölme yöntemini, gerçek hayattaki yüksek verimli teste en yakın yöntem olarak değerlendirmekteyiz.

Transfer-İlaç bölme, Soğuk-İlaç bölmeye benzer şekilde, test verilerinin kontaminasyonunu önlemek için tanıtılmıştır. İki bölme yöntemi arasındaki temel fark, bu Transfer-İlaç bölme yönteminin transfer öğrenme sırasında hücre hatları arasındaki kontaminasyonu da ortadan kaldırılmasıdır. Bu yöntem için Soğuk ilaç bölmede yapılan hedef model eğitiminde, kaynak modelin eğitiminde kullanılan bileşikler test katından çıkarılmıştır. Bu, test verilerindeki bileşiklerin, model için daha önce hiç görülmediğini garanti etmektedir. Çalışmanın devamında bu bölme yöntemine "transfer-ilaç-bölme" adı verilecektir.

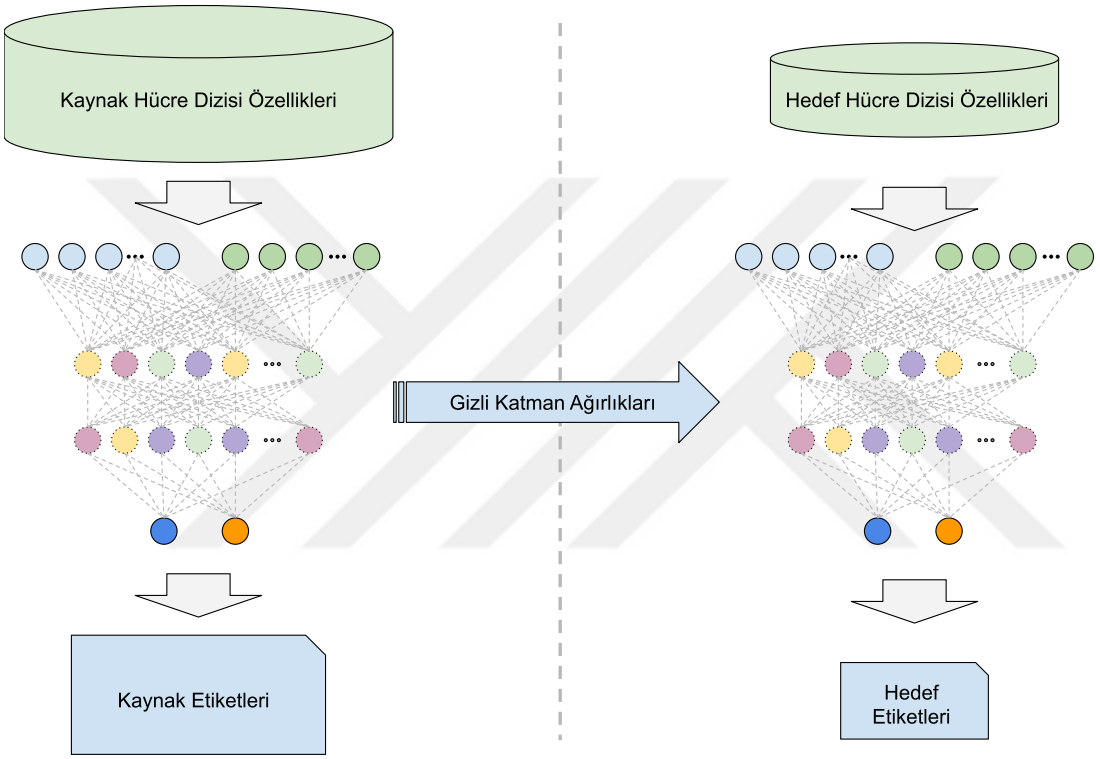


Şekil 4.1: Deneylerde kullanılan yapay sinir ağı.

4.3 Model Eğitimi

Bu çalışma, yeni test alanlarında önceki hücre dizilerinde test edilmiş bilgisini kullanmanın değerini doğrulamayı amaçlamaktadır. DeepCOP sonuçlarıyla doğrudan bir karşılaştırma yapabilmek için, eklediğimiz 2 veri bölme yöntemine ek olarak rastgele bölünmüş veri ile de deneylerimiz tekrarlanmaktadır. Bu deneyler sırasında, orijinal çalışmadan çok daha küçük sinir ağlarıyla benzer puanların elde edilebileceğinin sonucuna varılmıştır. Aşağıda listelenen deneylerin sonuçları, ağ yapısının geri kalanı değişmeden kalırken, orijinal 2 gizli katmanın nöron sayısının 400'e (DeepCOP'un 3155'inden) azaltıldığı bir yapay sinir ağı ile elde edilmiştir. Bu yapay sinir ağının görseli Şekil 4.1'de incelenebilir. Hücre dizilerinden öğrenilen bilgilerin diğer dizilere fayda sağlayabileceği varsayımına dayanarak, transfer öğrenme yöntemimiz olarak ağ tabanlı parametre paylaşımını kullanma kararı alınmıştır. Parametre paylaşım yöntemleri, ağın bir kısmını veya tamamını donör/kaynak problemi konusunda eğiterek problemler arasında öğrenilen bilgileri transfer etmektedir. Yapay ağ katmanlarını olduğu gibi almak veya eğitmek için yepyeni katmanlar ekleyerek aktarılan ağdaki hedef görev üzerinde eğitimi devam ettirmek bu yöntemler arasına girmektedir. Deneylerimizdeki transfer öğrenimi Şekil 4.2 ile görselleştirilmiştir.

Bu çalışmada, her bir hücre dizisi için %95 eğitim ve %5 doğrulama bölme verileriyle bir kaynak modeli eğittilmiştir. Daha sonra kaynak modelin nihai ağırlıklarını kaydettik ve bunları hedef hücre dizisi modelini eğitmek için başlangıç ağırlıkları



Şekil 4.2: Transfer öğrenimi: Daha büyük kaynak etki alanı verileri (ör. PC3) üzerinde eğitilen bir model, daha küçük hedef etki alanı verileri (ör. HEPG2) üzerinde eğitime devam eder.

olarak kullanılmıştır. Bu işlem bütün 6 kaynak ve 23 hedef hücre dizisi çiftleri için tekrarlanmıştır. Karşılaştırılabilirliği sağlamak için ağ yapısı ve parametreleri deneyler boyunca değişmeden tutulmuştur. DeepCOP çalışmasının orijinal altı hücre dizisi kaynak modellerimiz olarak seçilmiştir, çünkü bu altı hücreden sonra hücre dizisi başına düşen örnek sayısında büyük bir düşüş bulunmaktadır. Kaynak probleminde ne kadar fazla bilgi edinebilirse, hedef için daha güçlü bir temel oluşturulabilir.

5. DENEYLER

Karşılaştırma yapabilmek adına, ilk deneylerimiz DeepCOP'da kullanılan rastgele-ilaç-bölme üzerinde eğitilmiştir. Rastgele-ilaç modellerimizin sonuçları DeepCOP ile karşılaştıldığında, yukarı regüle edilmiş modellerin AUC puanları için ortalama %4,51 ve aşağı regüle edilmiş modellerin AUC puanları için ortalama %4,52 oranında iyileşme sağladığı gözlemlenmiştir. Yukarı regüle edilmiş modeller arasındaki maksimum iyileştirme, MCF7 ve NKDBA hücre dizisi arasında eğitilen modeldeydi; AUC puanında %15,73 iyileşme ile sonuçlandı. Aşağı regüleli modellere gelince, maksimum gelişme, A375 ile HL60 hücre dizisi arasındaydı; bu deney, AUC puanı için ek bir %12,73 ile sonuçlanmıştır.

Rastgele bölünmüş deneylerin ayrıntılı bir görünümü, yukarı ve aşağı regüle edilmiş model sonuçları Çizelge 5.1 ve Çizelge 5.2'da görülebilir. Her satır farklı hedef hücre satırlarını temsil etmektedir ve her sütun farklı kaynaklara karşılık gelirken, "TL Yok" (transfer öğrenme yok) sütunu, model kaynak model olmadan aynı bölmede eğitildiğinde ortaya çıkan modellerin sonuçlarını temsil etmektedir. Verilere bir bütün olarak bakıldığında net bir eğilim ortaya çıkmakta: En az iyileştirme, kaynak olarak seçtiğimiz hücre dizilerini hedef dizi olarak seçen deneylerde görülmüştür. Bu, her hücre dizisinin sahip olduğu örnek sayısıyla doğrudan ilişkilidir. Kalan 17 hedef hücre dizisi için, AUC puanları için ortalama kazançlar, yukarı ve aşağı regüle modelleri için %6.09 ve %6.04'tür. Skorlar etkileyici olsa da, rastgele bölme üçü arasında en az gerçekçi olanıdır. L1000 gibi transkriptomik veri kümeleri, ıslak laboratuvar deneylerinin gen bozulma ölçümlerinin alt kümeleridir. Bu deneylerde, bir hücre dizisinde kimyasal bir bileşik test edilir ve eş zamanlı olarak genlerin pertürbasyonları ölçülür. Bu nedenle, bir genin tepkilerini kullanarak başka bir genin pertürbasyonunu tahmin etmek, sadece o spesifik gen pertürbasyon değerinin kaybolduğu bir durumda faydalıdır. Bu, yüksek verimli taramayı simüle etmek için kullanışlı değildir.

Gerçekçi olmamasının yanı sıra, daha güçlü bir temel için ağ optimizasyonu ile DeepCOP'un izolasyon puanlarını iyileştirmeye çalışılırken bu, görünür bir sorun haline gelmiştir. Optimizasyon deneylerimiz, ağ yapısındaki büyük değişiklikler için bile değerlendirmede minimum değişiklik göstermiştir. Bir çok deney sonucunda elde edilen, neredeyse statik, ancak düşük hata oranının verilerin kendisiyle sınırlı olduğu sonucuna varılmıştır. Bu bölme yöntemiyle eğitilen yapay ağlar etkileyici sonuçlar üretirken, eğitim verisi test verisi ile kontamine olduğundan bu modeller veriyi istelilen

Çizelge 5.1: Rastgele-ilaç-bölme 10 kat verisi ile eğitilmiş modellerin; sütundan satıra transfer öğrenimi gerçeklenmiş yukarı regüle genlerin AUC puanları. Her satırdaki en iyi sonuç kalın harflerle belirtilmiştir.

	Kaynak Hücre						
Hedef Hücre ↓	**TL Yok**	**A375**	**A549**	**HT29**	**MCF7**	**PC3**	**VCAP**
A375	**0.8199**		0.8189	0.8190	0.8141	0.8167	0.8105
A549	**0.8254**	0.8243		0.8228	0.8246	0.8244	0.8222
BT20	0.7698	0.7975	**0.8015**	0.7994	0.8011	0.7987	0.7944
HA1E	0.8281	**0.8298**	0.8293	0.8281	0.8260	0.8297	0.8251
HCC515	0.8397	0.8405	0.8400	0.8403	0.8384	**0.8413**	0.8392
HEK293T	0.8011	0.8549	0.8533	0.8431	0.8503	**0.8576**	0.8495
HEPG2	0.8216	0.8548	0.8572	**0.8584**	0.8534	0.8517	0.8412
HL60	0.8427	**0.8964**	0.8836	0.8928	0.8872	0.8852	0.8835
HS578T	0.7556	0.7901	**0.7976**	0.7901	0.7909	0.7932	0.7859
HT29	0.7904	**0.7966**	0.7902		0.7875	0.7899	0.7822
HUH7	0.7505	0.8038	0.7998	**0.8072**	0.8048	0.8032	0.7987
JURKAT	0.8071	0.8579	0.8493	**0.8602**	0.8029	0.8536	0.8452
MCF10A	0.7733	**0.8054**	0.8002	0.8017	0.8013	0.7988	0.7928
MCF7	0.8313	0.8291	0.8315	0.8293		**0.8327**	0.8295
MDAMB231	0.7662	0.8207	0.8234	0.8214	0.8212	**0.8258**	0.8176
NKDBA	0.6833	0.7876	0.7891	0.7864	**0.7908**	0.7891	0.7878
NOMO1	0.7828	0.8013	**0.8278**	0.7485	0.7840	0.7816	0.7731
PC3	**0.8327**	0.8253	0.8276	0.8248	0.8279		0.8268
SKBR3	0.7655	0.7868	0.7943	0.7918	**0.7991**	0.7933	0.7855
THP1	0.7296	0.7492	0.7986	0.7805	0.7887	0.7883	**0.8130**
U266	0.7898	0.8146	0.8288	**0.8327**	0.7948	0.8307	0.8308
U937	0.7572	0.8023	0.8019	**0.8085**	0.7721	0.7993	0.7898
VCAP	**0.8471**	0.8457	0.8458	0.8450	0.8469	0.8466	

şekilde modelleyememekteydi. Daha küçük katmanlar benzer sonuçları daha çabuk üretebildiğinden, deneylerimizdeki katman boyutları değiştirilmiştir.

Bileşik tekrarını ortadan kaldırmak için soğuk-ilaç-bölünmesi tanıtılmıştır. Orijinal metodolojiyi bu yeni bölümle tekrarlamak, ortalama olarak %23,05 daha düşük AUC puanları ile sonuçlanmıştır. Bu sonuçlar Çizelge 5.3 ve Çizelge 5.4'ün TL Yok sütunlarında görülebilir. Yukarı regüleli modeller için en fazla gelişme, kaynak olarak A375 hücre dizisi verileri kullanıldığında JURKAT hücre dizisinde gözlemlenmiş; bu deneyde AUC puanı %22,81 artmıştır. Aşağı regüle için, PC3 kaynak olarak kullanıldığında %14,38 oranında iyileşme HCC515 hücre dizisinde gözlemlenmiştir. Soğuk-ilaç-bölme üzerindeki transfer öğrenme deney sonuçları Çizelge 5.3 ve Çizelge 5.4'te görülebilir. Daha önce de belirtildiği gibi, soğuk ilaç-bölme yönteminin üç bölme yöntemi arasındaki en gerçekçi yöntem olarak değerlendirildi. Bu bölme yöntemi üzerinde yapılan

Çizelge 5.2: Rastgele-ilaç-bölme 10 kat verisi ile eğitilmiş modellerin; sütundan satıra transfer öğrenimi gerçeklenmiş aşağı regüle genlerin AUC puanları. Her satırdaki en iyi sonuç kalın harflerle belirtilmiştir.

	Kaynak Hücre						
Hedef Hücre ↓	**TL Yok**	**A375**	**A549**	**HT29**	**MCF7**	**PC3**	**VCAP**
A375	0.8239		0.8244	**0.8248**	0.8214	0.8201	0.8165
A549	**0.8154**	0.8152		0.8132	0.8137	0.8153	0.8103
BT20	0.7758	0.8187	0.8173	0.8164	**0.8267**	0.8225	0.8102
HA1E	0.8384	**0.8403**	0.8380	0.8397	0.8382	0.8400	0.8342
HCC515	0.8420	0.8444	0.8440	0.8449	0.8448	**0.8473**	0.8423
HEK293T	0.8383	0.8835	0.8797	0.8766	0.8838	**0.8864**	0.8739
HEPG2	0.8341	0.8635	0.8638	0.8636	**0.8645**	0.8600	0.8507
HL60	0.7917	**0.8925**	0.8744	0.8881	0.8737	0.8739	0.8803
HS578T	0.7843	0.8080	**0.8082**	0.8037	0.8063	0.8077	0.7993
HT29	0.7974	**0.8032**	0.7954		0.7932	0.7960	0.7915
HUH7	0.7682	0.7954	0.7953	**0.7988**	0.7963	0.7957	0.7816
JURKAT	0.7917	**0.8781**	0.8506	0.8701	0.8591	0.8714	0.8422
MCF10A	0.7955	0.8200	0.8192	0.8196	**0.8215**	0.8165	0.8178
MCF7	**0.8430**	0.8425	0.8415	0.8414		0.8430	0.8395
MDAMB231	0.7953	0.8395	0.8367	0.8345	**0.8431**	0.8397	0.8335
NKDBA	0.7241	0.7948	0.7983	0.7914	**0.8002**	0.7967	0.7926
NOMO1	0.7934	0.8044	0.7665	0.7420	**0.8467**	0.8044	0.6887
PC3	0.8327	0.8322	0.8323	0.8314	**0.8345**		0.8320
SKBR3	0.7951	0.8233	0.8225	0.8255	**0.8292**	0.8203	0.8208
THP1	0.7590	0.7945	0.8156	0.7439	**0.8399**	0.7845	0.7082
U266	0.7765	0.8192	0.8076	0.8217	0.8344	0.8241	**0.8345**
U937	0.7569	**0.8050**	0.7642	0.8025	0.7370	0.7168	0.7705
VCAP	0.8564	0.8554	0.8541	0.8539	**0.8570**	0.8561	

deneyler, yukarı ve aşağı regüle edilmiş modeller için AUC puanlarında sırasıyla %9,70 ve %8,29'luk ortalama iyileştirmelerle sonuçlandı.

Tanıttığımız ikinci bölme yöntemi, transfer-ilaç-bölme, soğuk-ilaç-bölme ile karşılaştırıldığında biraz gerçekçi olmasa da, model eğitimi için önemli bir zorluk teşkil etmektedir. Kaynak hücre dizisinde eğitilen bileşikleri, hedef hücre dizisinin test kesitlerinden çıkartılarak soğuk ilaç ayırmada eğitilen modeller yeniden test edilmiştir. Transfer-ilaç-bölünmesi üzerinde, AUC puanları, yukarı ve aşağı regüle edilmiş modeller için ortalama %1,42 ve %-0,73 artmıştır. Bununla birlikte, maksimum AUC kazanımları diğer bölmelerdekilerle karşılaştırılabilir. Yukarı regüle için, JURKAT hücre dizisini A375'in üzerine eğitmek, %19,88 oranında iyileştirilmiş AUC ile sonuçlanmıştır. Aşağı regüleler arasındaki maksimum iyileşme, MCF7'den sonra eğitildiğinde AUC puanını %6,74 artıran HS578T hücre dizisidir. Transfer-ilaç-bölünmüş deneylerinin ayrıntılı

Çizelge 5.3: Soğuk-ilaç-bölme 10 kat verisi ile eğitilmiş modellerin; sütundan satıra transfer öğrenimi gerçeklenmiş yukarı regüle genlerin AUC puanları. Her satırdaki en iyi sonuç kalın harflerle belirtilmiştir.

	Kaynak Hücre						
Hedef Hücre ↓	**TL Yok**	**A375**	**A549**	**HT29**	**MCF7**	**PC3**	**VCAP**
A375	0.6145		0.6633	0.6499	0.6558	**0.6652**	0.6304
A549	0.5864	0.5944		0.5851	0.6294	**0.6445**	0.6145
BT20	0.5952	0.5460	0.6316	0.5801	**0.6553**	0.6268	0.6013
HA1E	0.6220	0.6601	0.6701	0.6482	0.6627	**0.6859**	0.6666
HCC515	0.6136	0.6401	0.6761	0.6507	0.6679	**0.6970**	0.6664
HEK293T	0.6165	**0.6971**	0.6238	0.6698	0.6543	0.6348	0.6466
HEPG2	0.6096	0.7297	0.7192	0.6761	0.6893	**0.7388**	0.6224
HL60	0.6063	0.6183	0.5397	**0.6665**	0.5670	0.5671	0.5604
HS578T	0.5762	0.5796	0.6509	0.6306	**0.6725**	0.6436	0.6307
HT29	0.6197	**0.6892**	0.6756		0.6742	0.6719	0.6374
HUH7	0.6428	0.6197	0.6856	0.6488	0.6749	**0.6980**	0.6679
JURKAT	0.5558	**0.6826**	0.5527	0.5816	0.5898	0.6307	0.6205
MCF10A	0.6434	0.6131	0.6360	0.6291	**0.6594**	0.6460	0.6323
MCF7	0.6059	0.6151	0.6569	0.6100		**0.6752**	0.6477
MDAMB231	0.6191	0.5859	0.6300	0.6138	**0.6896**	0.6586	0.6288
NKDBA	0.6121	0.5950	0.6267	0.6130	**0.6647**	0.6091	0.6129
NOMO1	0.5888	0.5100	0.5758	0.5689	0.5590	**0.5920**	0.4933
PC3	0.6055	0.6153	0.6453	0.6074	**0.6644**		0.6386
SKBR3	0.6057	0.5599	0.6161	0.5757	**0.6575**	0.6311	0.5835
THP1	0.5849	0.5427	0.5922	0.5841	**0.6019**	0.5848	0.5697
U266	**0.6430**	0.6557	0.6408	0.6136	0.6035	0.6249	0.5553
U937	0.5908	0.5800	0.6002	**0.6067**	0.5710	0.6046	0.5772
VCAP	0.5956	0.5970	0.6238	0.5924	0.6281	**0.6387**	

AUC puanları, yukarı ve aşağı regüle edilmiş model sonuçları sirasıyla Çizelge 5.5 ve Çizelge 5.6'da görülebilir.

Çizelge 5.4: Soğuk-ilaç-bölme 10 kat verisi ile eğitilmiş modellerin; sütundan satıra transfer öğrenimi gerçeklenmiş aşağı regüle genlerin AUC puanları. Her satırdaki en iyi sonuç kalın harflerle belirtilmiştir.

	Kaynak Hücre						
Hedef Hücre ↓	**TL Yok**	**A375**	**A549**	**HT29**	**MCF7**	**PC3**	**VCAP**
A375	0.6228		0.6648	0.6575	0.6563	**0.6671**	0.6317
A549	0.5799	0.5884		0.5933	0.6309	**0.6352**	0.6162
BT20	0.6236	0.6222	0.6620	0.6303	**0.7064**	0.6697	0.6188
HA1E	0.6331	0.6685	0.6673	0.6505	0.6693	**0.6954**	0.6723
HCC515	0.6184	0.6440	0.6836	0.6488	0.6885	**0.7073**	0.6776
HEK293T	0.6982	0.7408	0.7752	0.7292	**0.7781**	0.6995	0.7610
HEPG2	0.6653	**0.7457**	0.7450	0.7039	0.7415	0.7421	0.6554
HL60	**0.7314**	0.6375	0.6507	0.6811	0.6235	0.6327	0.6659
HS578T	0.6121	0.6203	0.6621	0.6469	**0.6964**	0.6575	0.6374
HT29	0.6199	**0.6969**	0.6709		0.6742	0.6718	0.6396
HUH7	0.6469	0.6543	0.6766	0.6777	**0.6955**	0.6708	0.6783
JURKAT	0.6679	**0.7213**	0.6766	0.6820	0.5887	0.6649	0.6359
MCF10A	0.6355	0.6396	0.6510	0.6559	**0.6805**	0.6800	0.6469
MCF7	0.6148	0.6240	0.6600	0.6207		**0.6950**	0.6575
MDAMB231	0.6595	0.6323	0.6702	0.6395	**0.7080**	0.6845	0.6370
NKDBA	0.6320	0.6387	0.6364	0.6289	**0.6673**	0.6345	0.6252
NOMO1	0.5725	0.5049	0.5858	0.5596	0.5649	**0.6268**	0.5099
PC3	0.6118	0.6166	0.6526	0.6209	**0.6740**		0.6489
SKBR3	0.6472	0.6325	0.6320	0.6474	**0.6937**	0.6545	0.6116
THP1	0.6326	0.5687	0.5944	0.5802	**0.6413**	0.6338	0.5806
U266	0.6799	0.6907	0.7053	0.6524	0.6506	**0.7065**	0.5438
U937	0.5641	0.5860	0.5903	0.5786	0.5697	0.5750	**0.5913**
VCAP	0.5965	0.6014	0.6197	0.6007	0.6272	**0.6458**	

Çizelge 5.5: Transfer-ilaç-bölme 10 kat verisi ile eğitilmiş modellerin; sütundan satıra transfer öğrenimi gerçeklenmiş yukarı regüle genlerin AUC puanları. Her satırdaki en iyi sonuç kalın harflerle belirtilmiştir.

	Kaynak Hücre						
Hedef Hücre ↓	**TL Yok**	**A375**	**A549**	**HT29**	**MCF7**	**PC3**	**VCAP**
A375	**0.6145**		0.6065	0.5943	0.5933	0.6060	0.5767
A549	**0.5864**	0.5414		0.5354	0.5711	0.5780	0.5572
BT20	0.5952	0.5156	0.5765	0.5440	**0.6019**	0.5729	0.5409
HA1E	0.6220	0.5979	0.6096	0.5913	0.6059	**0.6252**	0.6085
HCC515	0.6136	0.5843	0.6164	0.5899	0.6155	**0.6381**	0.6169
HEK293T	0.6165	**0.6302**	0.5973	0.5891	0.5685	0.5517	0.5673
HEPG2	0.6096	0.6504	0.6458	0.6255	0.6302	**0.6517**	0.5811
HL60	0.6063	0.5978	0.5479	**0.6268**	0.5572	0.5543	0.5590
HS578T	0.5762	0.5426	0.5796	0.5746	**0.6263**	0.5851	0.5655
HT29	0.6197	**0.6223**	0.6076		0.6028	0.5999	0.5781
HUH7	**0.6428**	0.5984	0.6334	0.6157	0.6274	0.6221	0.5978
JURKAT	0.5558	**0.6663**	0.6485	0.6325	0.6208	0.6652	0.6300
MCF10A	**0.6434**	0.5687	0.5845	0.5827	0.6160	0.6005	0.5722
MCF7	0.6059	0.5612	0.5896	0.5569		**0.6139**	0.5857
MDAMB231	0.6191	0.5545	0.5905	0.5852	**0.6397**	0.6074	0.5700
NKDBA	0.6121	0.5408	0.5623	0.5565	**0.6205**	0.5665	0.5797
NOMO1	**0.5888**	0.5472	0.4782	0.4862	0.4978	0.5132	0.4282
PC3	0.6055	0.5619	0.5874	0.5554	**0.6061**		0.5835
SKBR3	0.6057	0.5306	0.5651	0.5460	**0.6196**	0.5798	0.5416
THP1	**0.5849**	0.5095	0.5489	0.5360	0.5417	0.5522	0.4967
U266	0.6430	**0.6619**	0.6252	0.5920	0.5915	0.6056	0.5308
U937	**0.5968**	0.5570	0.5632	0.5792	0.5665	0.5739	0.5360
VCAP	**0.5956**	0.5446	0.5680	0.5419	0.5755	0.5845	

Çizelge 5.6: Transfer-ilaç-bölme 10 kat verisi ile eğitilmiş modellerin; sütundan satıra transfer öğrenimi gerçeklenmiş aşağı regüle genlerin AUC puanları. Her satırdaki en iyi sonuç kalın harflerle belirtilmiştir.

	Kaynak Hücre						
Hedef Hücre ↓	**TL Yok**	**A375**	**A549**	**HT29**	**MCF7**	**PC3**	**VCAP**
A375	**0.6228**		0.6085	0.6005	0.5962	0.6085	0.5765
A549	**0.5799**	0.5376		0.5403	0.5727	0.5719	0.5594
BT20	0.6236	0.5764	0.5992	0.5851	**0.6353**	0.6149	0.5696
HA1E	0.6331	0.6029	0.6081	0.5890	0.6085	**0.6337**	0.6155
HCC515	0.6184	0.5854	0.6166	0.5870	0.6239	**0.6455**	0.6270
HEK293T	**0.6982**	0.6496	0.5940	0.5941	0.6015	0.6014	0.6125
HEPG2	0.6653	**0.6693**	0.6693	0.6476	0.6639	0.6611	0.6072
HL60	**0.7314**	0.6355	0.6215	0.6722	0.6274	0.6164	0.6266
HS578T	0.6121	0.5725	0.6025	0.5900	**0.6534**	0.6029	0.5771
HT29	0.6199	**0.6274**	0.6050		0.6046	0.6039	0.5761
HUH7	**0.6469**	0.6113	0.6150	0.6275	0.6397	0.6126	0.6026
JURKAT	0.7107	**0.7162**	0.6795	0.6728	0.6581	0.6953	0.6088
MCF10A	**0.6355**	0.5980	0.6067	0.6163	0.6281	0.6200	0.6011
MCF7	0.6148	0.5686	0.5962	0.5646		**0.6200**	0.5938
MDAMB231	0.6595	0.5882	0.6082	0.6025	**0.6615**	0.6188	0.5855
NKDBA	**0.6320**	0.5843	0.5739	0.5785	0.6219	0.5668	0.5589
NOMO1	**0.5725**	0.4693	0.5309	0.4847	0.5083	0.5448	0.4489
PC3	0.6118	0.5634	0.5925	0.5655	**0.6156**		0.5940
SKBR3	0.6472	0.5858	0.5971	0.6041	**0.6495**	0.6131	0.5767
THP1	**0.6326**	0.5500	0.5382	0.5443	0.5967	0.5946	0.5165
U266	**0.6799**	0.6535	0.6438	0.5995	0.6150	0.6452	0.5164
U937	0.5641	0.5605	0.5575	0.5747	**0.5837**	0.5476	0.5779
VCAP	**0.5965**	0.5462	0.5637	0.5464	0.5753	0.5888	

6. TARTIŞMA

6.1 En İyi İyileştirmeler

Çizelge 6.1'daki en iyi 10 AUC puanına baktığımızda, yorum yapabileceğimiz birkaç eğilim gözlemleniyor.

Çizelge 6.1: En fazla iyileştirme gösteren transfer öğrenme deneylerinin diferansiyel ifade (DE) yönü ve bölme yöntemleri.

	Kaynak - Hedef	**İyileştirme (DE, Bölüm)**
1	A375 - JURKAT	%22,81 (Yukarı, soğuk)
2	PC3 - HEPG2	%21,19 (Yukarı, soğuk)
3	A375 - JURKAT	%19,88 (Yukarı, transfer)
4	A375 - HEPG2	%19,70 (Yukarı, soğuk)
5	A549 - JURKAT	%19,69 (Yukarı, transfer)
6	PC3 - JURKAT	%19,68 (Yukarı, transfer)
7	A549 - HEPG2	%17,98 (Yukarı, soğuk)
8	MCF7 - HS578T	%16,71 (Yukarı, soğuk)
9	A549 - JURKAT	%16,69 (Yukarı, transfer)
10	MCF7 - NKDBA	%15,73 (Yukarı, rastgele)

İlk olarak, tablodaki her deney bir yukarı diferansiyel ifade (DE) modelidir. Yukarı ve aşağı diferansiyel ifade deneylerinin ortalamasını ayrı ayrı alacak olursak, yukarı ayarlı deneylerin aşağı ayarlı deneylerden daha fazla etkilendiğini gözlemlenebilir. Bu durum, aşağı diferansiyele kıyasla daha yüksek yukarı diferansiyel örnek sayılarıyla ilişkilendirilebilir.

Görünen bir başka eğilim de bölme yöntemlerindedir. Bir istisna dışında, tablodaki her bölme yöntemi ya bir soğuk-ilaç-bölme ya da bir transfer-ilaç-bölmedir. Bunun başlıca nedeni, soğuk ilaç ayrımına geçildiğinde gözlemlenen AUC değerlerindeki %23'lük düşüş: Daha kötü bir temel üzerinde iyileştirme yapmak daha kolaydır. Ayrıca, sonraki bölümde tartışacağımız rastgele ilaç ayrımındaki komplikasyonlar, bu bölme yöntemindeki puanlarını iyileştirmeyi zorlaştırmaktadır.

Son olarak, iki hedef hücre hattı ilk 10'a hakimdir: JURKAT ve HEPG2. İlginç bir şekilde, bunların arkasındaki sebep birbirine tam olarak zıttır. HEPG2 için, transfer-ilaç-bölme deneylerinde benzer bir gelişme gözlemlenmemektedir, bu nedenle transfer öğrenmedeki iyileşmenin, kaynak hücre hattında test edilen aynı ilaçlardan gelmekte

olduğunu öngörebilir. JURKAT hücre hattı içinse, kaynak hücre hattında test edilen ilaçları ortadan kaldıran transfer-ilaç-bölmesine geçtiğimizde ortalama iyileşme %63 artmaktadır. Bu, hedef için daha iyi genellemeye ve kaynak hücre hattından aktarılan daha fazla bilgiye işaret etmektedir.

6.2 Rastgele Bölme Kullanmanın Komplikasyonları

DeepCOP'un arkasındaki ana fikir basit, ancak etkilidir: Çıktıların farklılaştırılabilir özelliklerini kullanmak, birden fazla genin pertürbasyon değerlerini tek bir hedef olarak ele alınmasını sağlamaktadır. Bu yöntem, her bileşik için veri miktarını arttırmıştır, bireysel genler arasındaki ilişkiyi korumuştur ve her çıktıya yeni bir anlam katmıştır. Ancak, bu teorinin gerçek hayattaki olayları simüle etmesi için popülasyonun gerekli doğrulama bölmesinden sonra yapılması gerekmektedir. Testlerimizde DeepCOP'a dahil edilen THP1 hücre hattı 18 farklı bileşik içermektedir. Orijinal verilerden, her bileşik için 978 gen bozulması elde edilebilir. Bu veriler çaprazlandıktan sonra 17,604 örnek elde edilir. Ancak elde sadece 18 özgün bileşik ve 978 özgün gen tanımlayıcı bulunmaktadır. Oluşan veri kümesini rastgele 10 kesite bölmek, veri sızıntısı olasılığını neredeyse garanti etmektedir.

Bu veri sızıntısının iki nedeni vardır; kopya genler ve bileşikler. L1000 veri kümesinde, bir deney bir bileşiğe ve bunun bir hücre hattı üzerindeki gen bozulma etkilerine karşılık gelmektedir. DeepCOP'un veri işleme adımından sonra, her bir geni ayrı ayrı tahmin etmek için 978 örnekte aynı bileşikle karşılaşılmaktadır. Bu verileri rastgele bölmek, tek bir deneyin gen çıktılarını farklı kesitlere ayırmak anlamına gelmektedir. Ayrıca bu, modelin gen bozulmalarını aynı bileşik için oluşan diğer genlerin sonuçlarından tahmin etmeyi öğrendiği anlamına gelmektedir. Bu gerçekçi olmayan ve önemsiz bir hedeftir, çünkü her gen pertürbasyonu o deney için aynı anda ölçülmektedir.

Gen tekrarından kaynaklanan veri sızıntısını önlemek için, 978 geni farklı kesitlere ve kesitlerdeki genleri bileşiklerle çaprazlayacağımız bir soğuk-gen bölmesi uygulamamız gerekir. Bununla birlikte, bu bölme yöntemi, HTS'nin amacının tam tersidir, çünkü HTS'de test edilmemiş bir bileşiğin mevcut genler üzerindeki etkilerini simüle etmeye çalışılmaktadır. Bu nedenle, bu bölme yöntemi bu çalışma için göz ardı edilebilir.

6.3 Kaynak Model Eğitiminde Rastgele Bölme

Bu çalışmada, kaynak modellerimizi erken durdurma ile rastgele %95-%5 eğitim doğrulama bölünmesi üzerinde eğitilmiştir. Rastgele bölme ile ilgili önceki açıklamala-

rımızdan, bu kaynak modellerin güvenilirliği de sorgulanmalıdır.

İlk deneylerimizde, kaynak modeller olarak rastgele bölme deneyleri için eğittiğimiz TL olmayan modelleri kullanmaktaydık. Rastgele bölme sorunu gün yüzüne çıktığında, kaynak modellerimiz kaynak verilerin 100'ünü kullanacak şekilde değiştirildi. Bu değişiklik, kaynak ağın böyle bir görev için optimize edilmemesi nedeniyle hedef model için negatif transfere neden olmuştur. Veri sızıntısının olumsuz etkilerini en aza indirgemek ve erken durdurmanın faydalarını korumak amacıyla; 95-%5 rastgele bölmeye karar verilmiştir. Unutulmamalıdır ki %95-%5 bölünmesi hala kaynak modellerin eğitimini olumsuz etkiler ve HTS'in yerini alacak bir model için; kaynak modeller, önce kimyasal tabanlı bir bölmede optimize edilmeli ve kaynak verilerin %100'ü üzerinde eğitilmelidir.

6.4 Yukarı ve Aşağı Düzenlemeleri Bölme

L1000'de gen bozulmaları 3 kategoriye ayrılabilir; yukarı regüle, nötr ve aşağı regüle. Çok sınıflı sınıflandırma, sınıflandırma problemlerinin özel bir durumudur ve genellikle ikili sınıflandırma problemlerine kıyasla modellemesi daha zordur.

DeepCOP ve bu çalışma normalde çok sınıflı bir sınıflandırma problemini 2 ikili sınıflandırmaya bölmektedir. Bu yaygın olarak kullanılan uygulama, verilerin modellenmesini kolaylaştırırken, aynı zamanda bilgi kaybına yol açmaktadır.

Spesifik olarak, bu uygulamada, yukarı ve aşağı regüle edilmiş çıktılar arasındaki ilişki kaybolur, çünkü ikili sınıflandırmalar "önemli ölçüde yukarı regüle edilmeyenler ve aktif olarak yukarı regüle edilenler" ve "önemli ölçüde aşağı regüle edilmeyenler ve aktif olarak aşağı regüle edilenler" olarak tanımlanır. İlk problem için bir optimal modeli eğitmek, aktif olarak aşağı-regüle-edilmiş bir örneği, aktif olarak yukarı-regüle-edilmiş olarak sınıflandırmak nötr bir örneği sınıflandırmaktan daha büyük bir ceza gerektirecektir. İkinci sınıflandırma problemi içinde tam tersi söylenebilir.

Problemi çok sınıflı olarak modellemek, veri setindeki özelliklerin en azından bu tür bir sinir ağı için, çok sınıflı modellemeden gelen zorluk ile başa çıkacak kadar açıklayıcı olmadığını göstermiştir. En belirgin çözüm, zıt olarak düzenlenmiş bir sorunun sonuçlarını kullanan özel bir kayıp işlevi kullanmak olacaktır. Başka bir olası çözüm, ek bir transfer öğrenimi olabilir. Aşağı regüle modellerde yukarı regüle edilmiş modellerin ağırlıklarını kullanmak veya aynı sinir ağının alt ağları olarak yukarı ve aşağı regülasyonu modellemek, kaybedilen bilgiyi kullanma potansiyeline sahiptir.

6.5 İkili Hale Getirme Eşiği

DeepCOP çalışmasında LINCS L1000 verisetinde bulunan rasyonel sayı halindeki çıktılar %5 ve %10 eşikleri kullanılarak ikili verilere dönüştürülür. Çalışma sonucunda %5 eşiğinde daha iyi sonuçlar alınmıştır bu tez çalışmasında da deneyler %5 eşiği kullanılarak gerçekleştirilmiştir. %5 eşiğinin %10 eşiğinden daha iyi sonuçlar vermesi daha düşük eşikler için daha yüksek sonuçlar alınabileceği varsayımını oluşturmuştur. Bu varsayımdan yola çıkılarak %2.5 eşiği ile ikililenmiş veri ile transfer öğrenme deneyleri tekrarlanmıştır. Bu deneylerin sonuçları Çizelge 6.2, 6.3, 6.4, ve 6.5'de görülebilir. Rastgele ilaç bölümünde ortalama %5.40, soğuk ilaç bölümünde yapılan deneylerde ise ortalama %9.58 iyileşme gözlemlenmiştir. %2.5 kullanılan deneylerde bazı hücre dizileri öğretilemez duruma geldiğinden (NOMO1) ve %5 eşiği ile yaptığımız deneyler arasında marjinal farklar olduğundan %5 eşikli deneylere önem verilmiştir.

Çizelge 6.2: %2.5 eşiği ile bölünmüş, rastgele-ilaç-bölme 10 kat verisi ile eğitilmiş modellerin; sütundan satıra transfer öğrenimi gerçeklenmiş yukarı regüle genlerin AUC puanları. Her satırdaki en iyi sonuç kalın harflerle belirtilmiştir.

	Kaynak Hücre						
Hedef Hücre ↓	**TL Yok**	**A375**	**A549**	**HT29**	**MCF7**	**PC3**	**VCAP**
A375	0.8445		**0.8527**	0.8504	0.8494	0.8525	0.8465
A549	0.8500	0.8499		0.8482	0.8521	**0.8534**	0.8506
BT20	0.7787	0.8221	0.8352	0.8304	**0.8366**	0.8317	0.8288
HA1E	0.8529	0.8627	0.8626	0.8594	0.8604	**0.8637**	0.8614
HCC515	0.8527	0.8663	0.8683	0.8666	0.8651	**0.8684**	0.8669
HEK293T	0.8392	0.8793	0.8853	**0.8962**	0.8879	0.8803	0.8706
HEPG2	0.8328	0.8723	**0.8751**	0.8697	0.8741	0.8729	0.8632
HL60	0.8583	0.8938	0.8993	0.8999	**0.9041**	0.9009	0.8986
HS578T	0.7665	0.8188	**0.8237**	0.8099	0.8163	0.8181	0.8151
HT29	0.8230	**0.8428**	0.8403		0.8409	0.8402	0.8361
HUH7	0.7696	0.8518	0.8558	0.8540	**0.8593**	0.8567	0.8517
JURKAT	0.7319	0.9100	0.9146	0.8823	0.9191	0.9205	**0.9232**
MCF10A	0.7807	**0.8329**	0.8273	0.8228	0.8302	0.8313	0.8232
MCF7	0.8697	0.8667	0.8664	0.8667		**0.8733**	0.8699
MDAMB231	0.7976	0.8619	0.8595	0.8550	0.8612	**0.8696**	0.8586
NKDBA	0.7586	**0.8328**	0.8253	0.8285	0.8270	0.8273	0.8281
NOMO1	**0.8215**	0.8082	0.7888	0.7613	0.7932	0.7674	0.7646
PC3	0.8717	0.8725	**0.8762**	0.8721	0.8724		0.8751
SKBR3	0.7749	0.8231	0.8265	0.8178	**0.8303**	0.8271	0.8177
THP1	0.7749	0.7750	0.7731	0.7349	0.8075	0.7733	**0.8121**
U266	0.8420	**0.8786**	0.9102	0.8476	0.8957	0.8799	0.9014
U937	0.7856	0.8502	0.8448	0.8528	**0.8557**	0.8500	0.8485
VCAP	**0.8746**	0.8722	0.8742	0.8709	0.8736	0.8745	

Çizelge 6.3: %2.5 eşiği ile bölünmüş, rastgele-ilaç-bölme 10 kat verisi ile eğitilmiş modellerin; sütundan satıra transfer öğrenimi gerçeklenmiş aşağı regüle genlerin AUC puanları. Her satırdaki en iyi sonuç kalın harflerle belirtilmiştir.

	Kaynak Hücre						
Hedef Hücre ↓	**TL Yok**	**A375**	**A549**	**HT29**	**MCF7**	**PC3**	**VCAP**
A375	0.8529		0.8626	0.8583	**0.8628**	0.8619	0.8577
A549	0.8460	0.8444		0.8408	0.8477	**0.8478**	0.8446
BT20	0.7877	0.8464	0.8531	0.8508	**0.8564**	0.8544	0.8488
HA1E	0.8634	0.8685	0.8716	0.8679	0.8689	**0.8728**	0.8689
HCC515	0.8583	0.8714	0.8740	0.8729	0.8734	**0.8766**	0.8731
HEK293T	**0.8713**	0.9142	0.9095	0.9042	0.9003	0.8871	0.8824
HEPG2	0.8513	0.8821	0.8830	0.8782	0.8858	**0.8870**	0.8792
HL60	0.8425	0.9028	**0.9058**	0.9023	0.8997	0.9022	0.9024
HS578T	0.8049	0.8503	0.8530	0.8456	0.8525	**0.8532**	0.8467
HT29	0.8293	0.8460	0.8458		0.8471	**0.8496**	0.8449
HUH7	0.8088	0.8449	0.8461	0.844	0.8473	0.8467	**0.8475**
JURKAT	0.7794	**0.9248**	0.9141	0.8520	0.8732	0.9140	0.8549
MCF10A	0.7999	0.8431	0.8423	0.8451	**0.8454**	0.8453	0.8396
MCF7	0.8804	0.8779	0.8809	0.8807		**0.8835**	0.881
MDAMB231	0.8245	0.8701	0.8719	0.8690	**0.8775**	0.8751	0.8704
NKDBA	0.7729	0.8489	0.8456	0.8477	0.8475	0.8472	**0.8521**
NOMO1	0.8329	0.8180	0.8243	0.8014	0.8336	0.8106	**0.8472**
PC3	0.8783	0.8775	0.8797	0.8763	**0.8821**		0.8796
SKBR3	0.8137	0.8477	0.8511	0.8509	**0.8603**	0.8537	0.8493
THP1	0.7698	0.7760	0.7935	0.7584	0.8191	**0.8424**	0.8205
U266	0.8309	0.8695	**0.8981**	0.8521	0.8786	0.8908	0.8620
U937	0.7798	**0.8279**	0.8163	0.8162	0.8046	0.8096	0.8132
VCAP	**0.8819**	0.8771	0.8802	0.8776	0.8787	0.8809	

Çizelge 6.4: %2.5 eşiği ile bölünmüş, soğuk-ilaç-bölme 10 kat verisi ile eğitilmiş modellerin; sütundan satıra transfer öğrenimi gerçeklenmiş yukarı regüle genlerin AUC puanları. Her satırdaki en iyi sonuç kalın harflerle belirtilmiştir.

	Kaynak Hücre						
Hedef Hücre ↓	**TL Yok**	**A375**	**A549**	**HT29**	**MCF7**	**PC3**	**VCAP**
A375	0.6778		0.7155	**0.7228**	0.7130	0.7183	0.6947
A549	0.6416	0.6468		0.6437	0.6830	**0.6913**	0.6613
BT20	0.6172	0.5954	0.6322	0.6104	**0.6583**	0.6468	0.6221
HA1E	0.6714	0.7202	0.7134	0.7076	0.7058	**0.7315**	0.7132
HCC515	0.6486	0.6913	0.7246	0.7068	0.7189	**0.7540**	0.7275
HEK293T	0.5942	0.7072	0.6257	0.6686	**0.7135**	0.6661	0.6301
HEPG2	0.6906	0.7660	0.7709	**0.7770**	0.7227	0.7638	0.7228
HL60	0.5965	0.6709	0.5914	**0.6782**	0.6173	0.5745	0.6499
HS578T	0.6167	0.6467	0.6515	0.6677	**0.6975**	0.6814	0.6700
HT29	0.6917	**0.7582**	0.7429		0.7351	0.7242	0.7083
HUH7	0.6633	0.7146	0.7126	0.7216	**0.7264**	0.7208	0.7048
JURKAT	0.6477	**0.7194**	0.5653	0.7165	0.6587	0.5854	0.6650
MCF10A	0.6866	0.6812	0.6709	0.6717	**0.7049**	0.6993	0.6536
MCF7	0.6648	0.6616	0.6956	0.6586		**0.7204**	0.6816
MDAMB231	0.6674	0.6513	0.6702	0.6688	**0.7018**	0.6950	0.6588
NKDBA	0.6029	0.6295	0.6402	0.6516	**0.6772**	0.6483	0.6522
PC3	0.6545	0.6570	0.6894	0.6571	**0.7008**		0.6830
SKBR3	0.6010	0.6134	0.6575	0.5955	**0.6789**	0.6735	0.6431
THP1	0.5612	0.5930	0.6220	0.6033	**0.6339**	0.5487	0.5560
U266	0.6154	0.6721	0.6737	0.6084	**0.7185**	0.5982	0.6613
U937	0.6374	**0.6644**	0.6067	0.6592	0.6374	0.6247	0.6415
VCAP	0.6510	0.6576	0.6855	0.6496	0.6847	**0.7119**	

Çizelge 6.5: %2.5 eşiği ile bölünmüş, soğuk-ilaç-bölme 10 kat verisi ile eğitilmiş modellerin; sütundan satıra transfer öğrenimi gerçeklenmiş aşağı regüle genlerin AUC puanları. Her satırdaki en iyi sonuç kalın harflerle belirtilmiştir.

	Kaynak Hücre						
Hedef Hücre ↓	**TL Yok**	**A375**	**A549**	**HT29**	**MCF7**	**PC3**	**VCAP**
A375	0.6857		0.7215	**0.7337**	0.7145	0.7182	0.7137
A549	0.6492	0.6584		0.6558	0.6883	**0.6971**	0.6719
BT20	0.6543	0.6408	0.6817	0.6828	**0.7225**	0.7007	0.6638
HA1E	0.6888	0.7316	0.7316	0.7154	0.7242	**0.7418**	0.7285
HCC515	0.6598	0.7056	0.7349	0.7095	0.7326	**0.7533**	0.7466
HEK293T	**0.8097**	0.8070	0.6807	0.7134	0.7975	0.7544	0.7266
HEPG2	0.7018	0.8071	0.7790	0.7975	0.7601	**0.8075**	0.7049
HL60	0.7420	0.7385	0.7427	**0.7794**	0.7523	0.6349	0.7603
HS578T	0.6434	0.6787	0.7031	0.6794	**0.7531**	0.6594	0.6654
HT29	0.6854	**0.7578**	0.7410		0.7407	0.7346	0.7155
HUH7	0.6853	0.7060	0.6991	**0.7222**	0.6994	0.7057	0.7044
JURKAT	0.6530	**0.7016**	0.6683	0.5775	0.6326	0.6937	0.6183
MCF10A	0.6951	0.6989	0.6780	0.7076	0.6974	0.7044	**0.7168**
MCF7	0.6716	0.6804	0.7050	0.6813		**0.7274**	0.6947
MDAMB231	0.6750	0.6851	0.6888	0.7011	**0.7413**	0.6916	0.6766
NKDBA	0.6862	0.6865	0.6883	**0.7035**	0.6896	0.6910	0.6749
NOMO1	0.6036	0.5111	0.5684	0.5596	**0.6185**	0.5745	0.5273
PC3	0.6649	0.6605	0.6907	0.6599	**0.7032**		0.69
SKBR3	0.6784	0.6463	0.6997	0.7064	**0.7316**	0.7108	0.6892
THP1	0.6287	0.6258	0.6169	0.6526	0.6881	**0.7208**	0.6569
U266	0.6137	0.6625	0.6803	0.5987	0.7122	**0.7324**	0.6100
U937	0.6300	0.6717	0.6244	**0.6768**	0.6427	0.6221	0.6647
VCAP	0.6240	0.6556	0.6853	0.6612	0.6856	**0.7109**	

7. SONUÇ

Bu tez çalışmasında meta-öğrenim yöntemlerinin -özellikle transfer öğreniminin-, yapay öğrenmedeki yeri ve önemi biyoenformatiğin alanının altında bulunan kemogenomik üzerinde deneysel bir şekilde gösterilmiştir. Yapılan çalışmada kenogenomiğe yeni bir özellik üretim şekli kazandıran ve bu yöntemi altı hücre hattı üzerinde test eden DeepCOP çalışması baz alınmıştır. DeepCOP çalışmasında yer alan metodoloji 23 hücre hattına genişletilmiş, çalışmanın başarısını daha doğru bir şekilde ölçen ve sonuçların daha iyi yorumlanmasını sağlayan iki veri bölme yöntemi eklenmiştir.

Yeni veri bölümünde yapılan testlerde DeepCOP çalışmasında alınan sonuçlarda %23,05'lik bir düşüş tespit edilmiş, orijinal veri bölümüyle karşılaştırabilmek adına bu testlerde eğitilen yapay öğrenim modellerinde orijinal veri bölümünde kullanılan parametreler değiştirilmemiştir. Çalışmada örnek sayısı az olan modellerde bilgi ve başarı arttırımı için, DeepCOP'da örnek çokluğu nedeni ile seçilen 6 hücre hattı üzerinde %95-%5 rastgele-ilaç veri bölüm yöntemi ile yukarı ve aşağı regüle olarak ayrılan toplamda 24 kaynak model üretilmiştir. Her kaynak modelden, kalan hücre hatlarında eğitilen modellere model-tabanlı transfer eğitimi yardımıyla bilgi aktarımı sağlanmıştır: Eğitilen kaynak hücre hatlarında eğitilen modellerin gizli katman ağırlıklar dondurularak hedef hücre dizilerinin başlangıç ağırlıkları olarak kullanılmıştır. 23 farklı hedef hücre dizisi üzerinde bu ağırlıklar ile rastgele-ilaç, soğuk-ilaç, ve transfer-ilaç veri bölme yöntemleri üzerinde 10-katlı çapraz doğrulama modelleri eğitilmiştir. DeepCOP'da kullanılan rastgele-ilaç bölme deneylerinin AUC skorlarında ortalama %4,52; gerçek ilaç deney ortamına en yakın olan soğuk-ilaç AUC skorlarında ortalama %9,00; hedef hücre hattındaki iyileşmeyi yorumlayabilmek için eklenen transfer-ilaç deneylerinin AUC skorlarında ise ortalama %0,345 iyileşme gözlemlenmiştir.

Bu tez çalışması, gelecekte yapılması planlanan çalışmalarda; kimyasalların gen regülasyonu tahmininde kullanılabilecek transfer öğrenimi yöntemlerine bir temel örnek olarak hazırlanmıştır. İlerleyen çalışmalarda ulaşılması hedeflerden ilki kaynak modellerin başarısını arttırmak olacaktır. Tartışma bölümünde belirttiğimiz gibi kaynak bölümlerinde kullanılan rastgele-ilaç bölme yöntemi modelin genelleştirilmesi önünde büyük bir engel. Bu modelin soğuk-ilaç veri bölme yöntemi üzerinde optimize edilmesi ve verinin %100'ünün kullanılarak eğitilmesi, hem kaynak hücre hattı testlerinde hem de hedef hücre hattı testlerinde optimal sonuçları verecektir. İlerleyen çalışmalarda denenecek ikinci bir hedef, temel bilgiler bölümünde bahsettiğimiz çeşitli transfer öğrenme tekniklerinin yanı sıra ikiden fazla hücre hattı modeli zincirlenmesidir.

KAYNAKLAR

[1] **Goodfellow**, **I.**, **Pouget-Abadie**, **J.**, **Mirza**, **M.**, **Xu**, **B.**, **Warde-Farley**, **D.**, **Ozair**, **S.**, **Courville**, **A. Bengio**, **Y.** (2014). Generative adversarial nets. *Advances in neural information processing systems* 27.

[2] **Wang**, **X.**, **Xie**, **L.**, **Dong**, **C. Shan**, **Y.** (2021). Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Sf. 1905–1914.

[3] **Maral**, **B. C.** (2022). Single Image Super-Resolution Methods: A Survey. *arXiv preprint arXiv:2202.11763*.

[4] **Bredel**, **M. Jacoby**, **E.** (2004). Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nature Reviews Genetics* 5.4, Sf. 262–275.

[5] **Hertzberg**, **R. P. Pope**, **A. J.** (2000). High-throughput screening: new technology for the 21st century. *Current opinion in chemical biology* 4.4, Sf. 445–451.

[6] **Bajorath**, **J.** (2002). Integration of virtual and high-throughput screening. *Nature Reviews Drug Discovery* 1.11, Sf. 882–894.

[7] **Vamathevan**, **J.**, **Clark**, **D.**, **Czodrowski**, **P.**, **Dunham**, **I.**, **Ferran**, **E.**, **Lee**, **G.**, **Li**, **B.**, **Madabhushi**, **A.**, **Shah**, **P.**, **Spitzer**, **M.**, et al. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery* 18.6, Sf. 463–477.

[8] **Fourches**, **D.**, **Muratov**, **E. Tropsha**, **A.** (2015). Curation of chemogenomics data. *Nature chemical biology* 11.8, Sf. 535–535.

[9] **Subramanian**, **A.**, **Narayan**, **R.**, **Corsello**, **S. M.**, **Peck**, **D. D.**, **Natoli**, **T. E.**, **Lu**, **X.**, **Gould**, **J.**, **Davis**, **J. F.**, **Tubelli**, **A. A.**, **Asiedu**, **J. K.**, et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171.6, Sf. 1437–1452.

[10] **Pal**, **S. Mitra**, **S.** (1992). Multilayer perceptron, fuzzy sets, and classification. *IEEE Transactions on Neural Networks* 3.5, Sf. 683–697.

[11] **Dayhoff**, **J. E.** (1990). Neural network architectures: an introduction. Van Nostrand Reinhold Co.

[12] **Rumelhart**, **D. E.**, **Hinton**, **G. E. Williams**, **R. J.** (1985). *Learning internal representations by error propagation*. Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science.

[13] **Radford**, **A.**, **Metz**, **L. Chintala**, **S.** (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

[14] **Lamb**, **J.**, **Crawford**, **E. D.**, **Peck**, **D.**, **Modell**, **J. W.**, **Blat**, **I. C.**, **Wrobel**, **M. J.**, **Lerner**, **J.**, **Brunet**, **J.-P.**, **Subramanian**, **A.**, **Ross**, **K. N.**, et al.

(2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *science* 313.5795, Sf. 1929–1935.

[15] **Gobbi**, **A. Poppinger**, **D.** (1998). Genetic optimization of combinatorial libraries. *Biotechnology and bioengineering* 61.1, Sf. 47–54.

[16] **Rogers**, **D. Hahn**, **M.** (2010). Extended-connectivity fingerprints. *Journal of chemical information and modeling* 50.5, Sf. 742–754.

[17] **Consortium**, **G. O.** (2015). Gene ontology consortium: going forward. *Nucleic acids research* 43.D1, Sf. D1049–D1056.

[18] **Wang**, **Z.**, **Clark**, **N. R. Ma'ayan**, **A.** (2016). Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics* 32.15, Sf. 2338–2345.

[19] **Chen**, **Y.**, **Li**, **Y.**, **Narayan**, **R.**, **Subramanian**, **A. Xie**, **X.** (2016). Gene expression inference with deep learning. *Bioinformatics* 32.12, Sf. 1832–1839.

[20] **Cobanoglu**, **M. C.**, **Liu**, **C.**, **Hu**, **F.**, **Oltvai**, **Z. N. Bahar**, **I.** (2013). Predicting drug–target interactions using probabilistic matrix factorization. *Journal of chemical information and modeling* 53.12, Sf. 3399–3409.

[21] **Kidd**, **B. A.**, **Wroblewska**, **A.**, **Boland**, **M. R.**, **Agudo**, **J.**, **Merad**, **M.**, **Tatonetti**, **N. P.**, **Brown**, **B. D. Dudley**, **J. T.** (2016). Mapping the effects of drugs on the immune system. *Nature biotechnology* 34.1, Sf. 47–54.

[22] **Mayr**, **A.**, **Klambauer**, **G.**, **Unterthiner**, **T.**, **Steijaert**, **M.**, **Wegner**, **J. K.**, **Ceulemans**, **H.**, **Clevert**, **D.-A. Hochreiter**, **S.** (2018). Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical science* 9.24, Sf. 5441–5451.

[23] **Chen**, **H.**, **Cheng**, **F. Li**, **J.** (2020). iDrug: Integration of drug repositioning and drug-target prediction via cross-network embedding. *PLoS computational biology* 16.7, e1008040.

[24] **Bynagari**, **N. B.** (2018). On the ChEMBL Platform, a Large-scale Evaluation of Machine Learning Algorithms for Drug Target Prediction. *Asian Journal of Applied Science and Engineering* 7, Sf. 53–64.

[25] **Zhou**, **M.**, **Chen**, **Y. Xu**, **R.** (2019). A drug-side effect context-sensitive network approach for drug target prediction. *Bioinformatics* 35.12, Sf. 2100–2107.

[26] **Wen**, **M.**, **Zhang**, **Z.**, **Niu**, **S.**, **Sha**, **H.**, **Yang**, **R.**, **Yun**, **Y. Lu**, **H.** (2017). Deep-learning-based drug–target interaction prediction. *Journal of proteome research* 16.4, Sf. 1401–1409.

[27] **Öztürk**, **H.**, **Özgür**, **A. Ozkirimli**, **E.** (2018). DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* 34.17, Sf. i821–i829.

[28] **Nickel**, **J.**, **Gohlke**, **B.-O.**, **Erehman**, **J.**, **Banerjee**, **P.**, **Rong**, **W. W.**, **Goede**, **A.**, **Dunkel**, **M. Preissner**, **R.** (2014). SuperPred: update on drug classification and target prediction. *Nucleic acids research* 42.W1, W26–W31.

[29] **Pliakos**, **K. Vens**, **C.** (2020). Drug-target interaction prediction with tree-ensemble learning and output space reconstruction. *BMC bioinformatics* 21.1, Sf. 1–11.

[30] **Chen**, **X.**, **Yan**, **C. C.**, **Zhang**, **X.**, **Zhang**, **X.**, **Dai**, **F.**, **Yin**, **J. Zhang**, **Y.** (2016). Drug–target interaction prediction: databases, web servers and computational models. *Briefings in bioinformatics* 17.4, Sf. 696–712.

[31] **Wishart**, **D. S.**, **Feunang**, **Y. D.**, **Guo**, **A. C.**, **Lo**, **E. J.**, **Marcu**, **A.**, **Grant**, **J. R.**, **Sajed**, **T.**, **Johnson**, **D.**, **Li**, **C.**, **Sayeeda**, **Z.**, et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research* 46.D1, Sf. D1074–D1082.

[32] **Gaulton**, **A.**, **Bellis**, **L. J.**, **Bento**, **A. P.**, **Chambers**, **J.**, **Davies**, **M.**, **Hersey**, **A.**, **Light**, **Y.**, **McGlinchey**, **S.**, **Michalovich**, **D.**, **Al-Lazikani**, **B.**, et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* 40.D1, Sf. D1100–D1107.

[33] **Kuhn**, **M.**, **Letunic**, **I.**, **Jensen**, **L. J. Bork**, **P.** (2016). The SIDER database of drugs and side effects. *Nucleic acids research* 44.D1, Sf. D1075–D1079.

[34] **Tang**, **J.**, **Szwajda**, **A.**, **Shakyawar**, **S.**, **Xu**, **T.**, **Hintsanen**, **P.**, **Wennerberg**, **K. Aittokallio**, **T.** (2014). Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling* 54.3, Sf. 735–743.

[35] **Turki**, **T.**, **Wei**, **Z. Wang**, **J. T.** (2017). Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. *IEEE Access* 5, Sf. 7381–7393.

[36] — (2018). A transfer learning approach via procrustes analysis and mean shift for cancer drug sensitivity prediction. *Journal of bioinformatics and computational biology* 16.03, Sf. 1840014.

[37] **Dhruba**, **S. R.**, **Rahman**, **R.**, **Matlock**, **K.**, **Ghosh**, **S. Pal**, **R.** (2018). Application of transfer learning for cancer drug sensitivity prediction. *BMC bioinformatics* 19.17, Sf. 51–63.

[38] **Zhu**, **Y.**, **Brettin**, **T.**, **Evrard**, **Y. A.**, **Partin**, **A.**, **Xia**, **F.**, **Shukla**, **M.**, **Yoo**, **H.**, **Doroshow**, **J. H. Stevens**, **R. L.** (2020). Ensemble transfer learning for the prediction of anti-cancer drug response. *Scientific reports* 10.1, Sf. 1–11.

[39] **Qiang**, **B.**, **Lai**, **J.**, **Jin**, **H.**, **Zhang**, **L. Liu**, **Z.** (2021). Target prediction model for natural products using transfer learning. *International journal of molecular sciences* 22.9, Sf. 4632.

[40] **Ke**, **G.**, **Meng**, **Q.**, **Finley**, **T.**, **Wang**, **T.**, **Chen**, **W.**, **Ma**, **W.**, **Ye**, **Q. Liu**, **T.-Y.** (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30.

[41] **Yang**, **W.**, **Soares**, **J.**, **Greninger**, **P.**, **Edelman**, **E. J.**, **Lightfoot**, **H.**, **Forbes**, **S.**, **Bindal**, **N.**, **Beare**, **D.**, **Smith**, **J. A.**, **Thompson**, **I. R.**, et al. (2012). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research* 41.D1, Sf. D955–D961.

[42] **Barretina**, **J.**, **Caponigro**, **G.**, **Stransky**, **N.**, **Venkatesan**, **K.**, **Margolin**, **A. A.**, **Kim**, **S.**, **Wilson**, **C. J.**, **Lehár**, **J.**, **Kryukov**, **G. V.**, **Sonkin**, **D.**, et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483.7391, Sf. 603–607.

[43] **Chawla**, **N. V.**, **Bowyer**, **K. W.**, **Hall**, **L. O. Kegelmeyer**, **W. P.** (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16, Sf. 321–357.

[44] **Smola**, **A. J. Schölkopf**, **B.** (2004). A tutorial on support vector regression. *Statistics and computing* 14.3, Sf. 199–222.

[45] **Boser**, **B. E.**, **Guyon**, **I. M. Vapnik**, **V. N.** (1992). A Training Algorithm for Optimal Margin Classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. COLT '92. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, Sf. 144–152. URL: https://doi.org/10.1145/130385.130401.

[46] **Hoerl**, **A. E. Kennard**, **R. W.** (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12.1, Sf. 55–67.

[47] **Hosmer Jr**, **D. W.**, **Lemeshow**, **S. Sturdivant**, **R. X.** (2013). Applied logistic regression. Vol. 398. John Wiley & Sons.

[48] **Mei**, **S.**, **Fei**, **W. Zhou**, **S.** (2011). Gene ontology based transfer learning for protein subcellular localization. *BMC bioinformatics* 12.1, Sf. 1–12.

[49] **Mignone**, **P.**, **Pio**, **G.**, **D'Elia**, **D. Ceci**, **M.** (Oct. 2019). Exploiting transfer learning for the reconstruction of the human gene regulatory network. *Bioinformatics* 36.5, Sf. 1553–1561. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/5/1553/32793797/btz781.pdf. URL: https://doi.org/10.1093/bioinformatics/btz781.

[50] **O'Donovan**, **S. D.**, **Driessens**, **K.**, **Lopatta**, **D.**, **Wimmenauer**, **F.**, **Lukas**, **A.**, **Neeven**, **J.**, **Stumm**, **T.**, **Smirnov**, **E.**, **Lenz**, **M.**, **Ertaylan**, **G.**, et al. (2020). Use of deep learning methods to translate drug-induced gene expression changes from rat to human primary hepatocytes. *PloS one* 15.8, e0236392.

[51] RDKit: Open-source cheminformatics (n.d.). [Online; accessed 15-April-2020].