**T.C.**
**SÜLEYMAN DEMİREL ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# ALZHEİMER MODELİNDE: İLAÇ-HEDEF ETKİLEŞİMİ TAHMİNİ

**Münevver DEMİR**

**Danışman**
**Prof. Dr. Selçuk ÇÖMLEKÇİ**

**II. Danışman**
**Prof. Dr. Ecir Uğur KÜÇÜKSİLLE**

**YÜKSEK LİSANS TEZİ**
**BİYOMÜHENDİSLİK ANABİLİM DALI**
**ISPARTA- 2023**

## İÇİNDEKİLER

**Sayfa**

# ÖZET

**Yüksek Lisans Tezi**

**ALZHEİMER MODELİNDE: İLAÇ-HEDEF ETKİLEŞİMİ TAHMİNİ**

**Münevver DEMİR**

**Süleyman Demirel Üniversitesi**
**Fen Bilimleri Enstitüsü**
**Biyomühendislik Anabilim Dalı**

**Danışman: Prof. Dr. Selçuk ÇÖMLEKÇİ**

**II. Danışman: Prof. Dr. Ecir Uğur KÜÇÜKSİLLE**

İlaçlar, bir hedefin biyolojik işlevini aktive etmek veya inhibe etmek için hedef proteinlerle etkileşime girerek çalışır. Moleküler terapötik hedeflerin uygun şekilde tanımlanması, bir hastalık için güvenli ve etkili ilaçların geliştirmesinde çok önemlidir. İlaç geliştirme için hedeflenecek yeni proteinlerin tanımlanmasındaki ve yeni terapötik adayların belirlenmesindeki kritik önemi nedeniyle, ilaç-hedef etkileşimlerinin (İHE) tahmini, son zamanlarda önemli bir araştırma faaliyeti alanı olarak öne çıkmıştır. İHE tahmini için kullanılan geleneksel yöntemler, zaman alıcı ve pahalıdır. Bu sebeple İHE tahmininde in siliko yöntemlerin geliştirilmesine ihtiyaç vardır. Hesaplamalı yöntemlerin İHE'de doğru tahminleri, in vitro uygulamalardaki araştırma alanını azaltacak ve birçok hastalığın tedavi sürecini geliştirecektir. Özellikle bu sürecin Alzheimer gibi, hastalığın seyrini değiştiren ilaçların bulunmadığı rahatsızlıklarda ayrı bir önemi vardır.

Alzheimer hastalığı(AH), sinir sisteminin hasar görmesi ve sinir hücrelerinin ölmesi sonucuyla oluşan nörodejeneratif bir rahatsızlıktır. AH; kademeli hafıza gerilemelerine, işlevsel engelliliğe, hafıza kaybına neden olan ve kesin bir tedavisi bulunmayan ciddi bir hastalıktır. Bu tez çalışmasında,İHE tahmini yapılmak üzere AH'na ait bilinen protein hedef verileri ve ilgili proteinlere ait etkileşimi bilinen ilaç verileri kullanılmıştır.

Protein temsili için AH'na ait proteinlerin aminoasit dizilerinden protein vektörleri (ProtVec) elde edilmiştir. İlaç temsili için ilaçların SMILES dizilerinden elde edilen iki boyutlu moleküler yapı görüntülerinin ayrı ayrı işlenmek üzere sırasıyla evrişimsel sinir ağları(CNN) ve transformers ağları kullanılarak öznitelikleri çıkarılmıştır. İlaç hedef etkileşiminde etkileşimi bilinmeyen (negatif örnekler) çiftler; birbirlerine en uzak protein vektörleri arasından rastgele seçilmesinin yanında; Öklid, Minkowski, Manhattan yöntemleri ile elde edilmiş ve ilgili proteine ait ilaçlar etiketlenerek oluşturulmuştur. İHE tahmini için veriler, sonuçları birbirleriyle karşılaştırılmak üzere ilaç verileri için iki farklı yöntem kullanılarak özniteliklerinin çıkarıldığı temsiller ile ayrı ayrı birleştirilmiştir. Temsilleri oluşturulan ilaç, hedef, etkileşim verileri birleştirilmiş ve makine öğrenimi sınıflandırıcı algoritmaları ile modellenmiştir. Modelleme aşamasında; Rastgele Orman (RO), Lojistik Regresyon (LR), Karar Ağacı (KA) ve Destek Vektör Makinesi (DVM) sınıflandırma algoritmaları kullanılmıştır.

Model performansları; Area Under the Curve (AUC), Area Under the Precision-Recall Curve (AUPRC), f1 puanı, Matthew's Correlation Coefficient (MCC) ve doğruluk metrikleri kullanılarak değerlendirilmiştir. Değerlendirme sonuçları karşılaştırıldığında, RO modeline ait ilaç temsili için transformers ağı kullanılan veri setinin AUC 88.98(%) ve AUPR 94.24(%) metrik değerleri ile en iyi test sonuçlarını verdiği gözlemlenmiştir. İHE etkisini iyileştirmek üzere uygulan modelin, ilaç yeniden konumlandırma alanına faydalı olabilmesi adına moleküler kenetleme(docking) ve ileri çalışmalar ile güçlendirilmesine ihtiyacı vardır.

**Anahtar Kelimeler:** Alzheimer, ilaç-hedef etkileşim tahmini, ilaç keşfi, python

**2023, 54 sayfa**

## ABSTRACT

**M.Sc. Thesis**

**DRUG-TARGET INTERACTION PREDICTION IN THE ALZHEIMER MODEL**

**Münevver DEMİR**

**Süleyman Demirel University**
**Graduate School of Natural and Applied Sciences**
**Department of Bioengineering**

**Supervisor: Prof. Dr. Selçuk ÇÖMLEKÇİ**

**Co-Supervisor: Prof. Dr. Ecir Uğur KÜÇÜKSİLLE**

Drugs work by interacting with target proteins to activate or inhibit the biological function of a target. Appropriate identification of molecular therapeutic targets is crucial in the development of safe and effective drugs for a disease. Due to its critical importance in identifying novel proteins to target for drug development and in identifying new therapeutic candidates, the prediction of drug-target interactions (DTIs) has recently emerged as an important area of research activity. Traditional methods used for DTI prediction are time-consuming and expensive. Therefore, there is a need to develop in silico methods for DTI prediction. Accurate prediction of DTI by computational methods will reduce the need for in vitro applications and improve the treatment process of many diseases. This is particularly important in disorders such as Alzheimer's disease, where there are no drugs that alter the course of the disease.

Alzheimer's disease (AD) is a neurodegenerative disorder caused by damage to the nervous system and the death of nerve cells. AD is a serious disease that causes gradual memory decline, functional disability, memory loss, and has no definitive cure. In this thesis, known protein target data of AD and drug data with known interactions with related proteins were used to predict DTIs.

For protein representation, protein vectors (ProtVec) were obtained from the amino acid sequences of AD proteins. For drug representation, two-dimensional molecular structure images of the drugs obtained from SMILES arrays were processed separately, and features were extracted using convolutional neural networks (CNN) and transformer networks, respectively. In drug-target interaction, pairs with unknown interaction (negative samples) were randomly selected from the most distant protein vectors, obtained by Euclidean, Minkowski, Manhattan methods, and labeled with the drugs belonging to the relevant protein. For DTI prediction, the data were combined separately with the representations from which the features were extracted by two different methods for the drug data to compare the results with each other. The drug, target, and interaction data were combined and modeled with machine learning classifier algorithms. Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), and Support Vector Machine (SVM) classification algorithms were used in the modeling phase. Model performances were evaluated using Area Under the Curve

(AUC), Area Under the Precision-Recall Curve (AUPRC), F1 score, Matthew's Correlation Coefficient (MCC), and accuracy metrics.

Comparing the evaluation results, it was observed that the dataset using the transformer network for drug representation of the RF model achieved the best test results of AUC 88.98% and AUPR 94.24%. The model applied to improve the DTI effect needs to be strengthened with molecular docking and future studies to be useful in the field of drug repositioning.

**Keywords:** alzheimer, drug-target interaction prediction, drug discovery, python

**2023, 54 pages**

**TEŞEKKÜR**

Bu araştırma için beni yönlendiren, karşılaştığım zorlukları bilgi ve tecrübesi ile aşmamda yardımcı olan, bana her daim yardımcı olan çok değerli Danışman Hocalarım Prof. Dr. Ecir Uğur KÜÇÜKSİLLE ve Prof. Dr. Selçuk ÇÖMLEKÇİ' ye teşekkürlerimi sunarım. Literatür araştırmalarımda her zaman bana yardımcı olan ve her daim beni destekleyen çok değerli hocam Dr. M. Emrah ŞELLİ ve ailesine çok teşekkür ederim.

Araştırmanın yürütülmesinde her zaman bana destek olan canım arkadaşlarım Ali ÖZTÜRK'e, Ceyhan ÜNLÜ'ye, Ayşenur ÇETİNKAYA'ya, Serra KERVANOĞLU'na, Serra Nur GÜNNÜ'ye, Şeyma ÖZLEN'e, Merve MAŞA'ya ve Ali Yavuz ÇAKIR'a çok teşekkür ederim.

Tezimin her aşamasında beni yalnız bırakmayan, her zaman beni canı gönülden destekleyen, her yorulduğumda bana destek olan canım eşime ve çok kıymetli anneme, babama ve kardeşime sonsuz sevgi ve saygılarımı sunarım.

Münevver DEMİR
ISPARTA, 2023

## ŞEKİLLER DİZİNİ

**Sayfa**

## ÇİZELGELER DİZİNİ

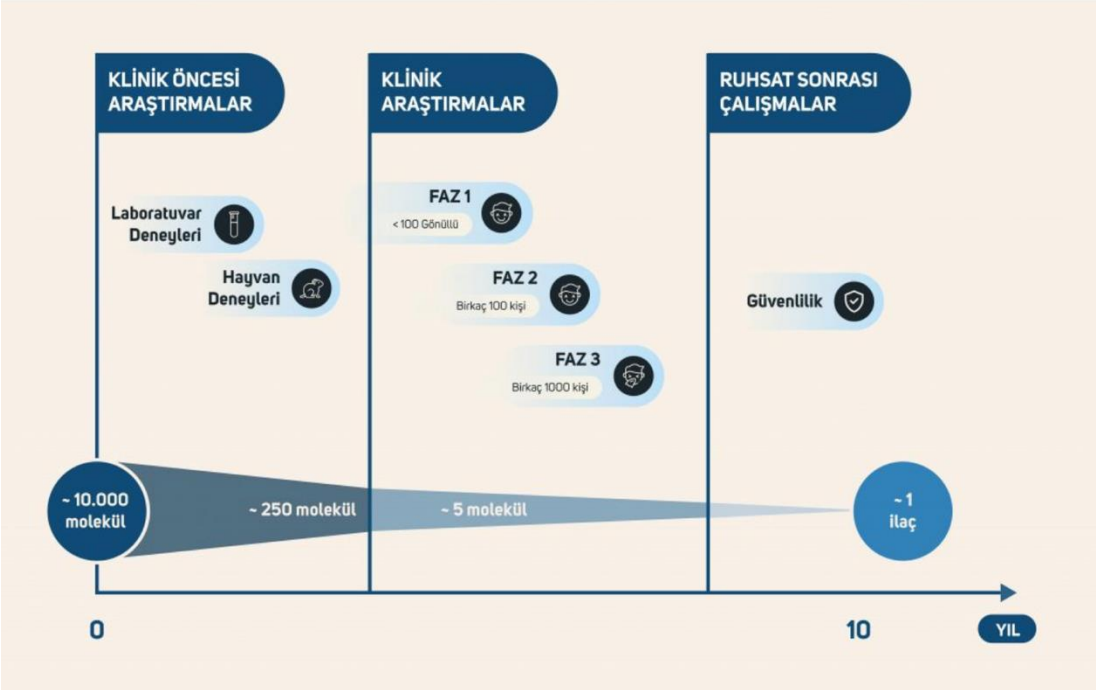
**Sayfa**

## KISALTMALAR DİZİNİ

| | |
|---|---|
| APP | Amiloid Precursor Proteini |
| BPE | Bileşik Protein Etkileşimleri |
| DVM | Destek Vektör Makineleri |
| İHE | İlaç hedef etkileşimi |
| KA | Karar Ağaçları |
| LR | Lojistik Regresyon |
| MCC | Matthews Correlation Coefficient |
| RO | Rastgele Orman |
| SMILES | Simplified-Molecular-Input-Line-Entry-System |

# 1. GİRİŞ

Hızla gelişen dünyamızda, her geçen gün hastalıklar artmakta ve evrilmektedir. Hastalıkların her daim insan hayatında olduğu gerçeği ise, ilaçların önemi ve gerekliliği açısından yadsınamaz şekilde ciddiyetini korumaktadır. İlaç sanayisinin gelişme hızının, yeni hastalıkların oluşma hızına olan doğru orantısı tedavi süreçlerini kısaltacaktır. Ayrıca, hastalıkların etkili ve doğru tedaviler ile çaresinin bulunması da toplumun ilerleme kaydetmesi açısından çok önemlidir (Kim vd., 2021; Sachdev ve Gupta, 2019). İlaç geliştirmede kullanılan geleneksel yöntemlerin masrafının yüksekliği, ilaç sektöründe büyük bir problemdir ve bu sürecin sonunda yüksek başarı oranının kesinlikle garantisi yoktur. Bir ilacın geliştirilme sürecine ait görsel Şekil 1.1.' de gösterilmiştir (Raghavendra vd., 2012). 15 yılı bulan çalışmalar neticesinde yeni bir ilaç oluşturmanın maliyeti tahminlere göre yaklaşık 2,6 milyar ABD dolarını bulabilmektedir (Doytchinova, 2022). Bu sebeple, hem maliyetin düşürülmesi hem de sürecin kısaltılması açısından ilaç hedef etkileşimi (İHE) tahmini yöntemleri gibi yeni yöntemlerin geliştirilmesine ihtiyaç bulunmaktadır.




Şekil 1.1. İlaç keşif süreci (Raghavendra vd., 2012).

İlaç hedef etkileşimi (İHE) tahmini, ilaç keşfinin ilk ve en önemli aşamasıdır. Hesaplamalı yöntemler ile doğru tahminin yapılması ilaç keşif sürecini kısaltmakla birlikte maliyetleri de azaltacaktır. Ayrıca İHE tahmini ile yapılan ilaç keşfinin, potansiyel ilaçların laboratuvardaki araştırma alanını daraltması ve hata oranını minimum seviyede tutması beklenmektedir.(Kim vd., 2021; Nanor vd., 2020). İlaç keşfinin en önemli adımlarından biri olan ilaçların ve proteinlerin arasındaki ilişkinin tanımlanması doğru tedavi için kritik bir alan olsa da hala günümüzde bilinmezliğin yoğun olduğu alanlardan birisidir. Diğer bir yönden, mevcut İHE tahmin tekniklerinin sınırlı hassasiyeti ve yüksek yanlış pozitif oranı, performanslarını olumsuz yönde etkilemektedir. Kimyasal ve genomik araştırma alanlarının arasındaki bilgilerin sınırlı olması, İHE'lerinin verimli şekilde tespit edebilecek yeni yöntemlerin geliştirilmesinin arkasındaki itici güçtür (Yamanishi vd., 2008).


Şekil 1.2. İlaç hedef etkileşimi süreci (Sachdev ve Gupta, 2019).

Şekil 1.2.'de İHE süreci özetlemektedir (Sachdev ve Gupta, 2019). Şekil 1.2.' de görüldüğü üzere, öncelikle ilaç ilgili proteine geçici bir bağ oluşturarak bağlanır. Bağlanan ilaç pozitif veya negatif bir biyolojik değişiklik oluşturmak üzere hedef protein ile reaksiyona girer ve sonrasında hedefi terk eder. Dolayısıyla bu süreç vücutta da aynı doğrultuda işlemektedir. İHE, protein hedefleri ile kimyasal bileşikler arasındaki etkileşimlerin tanımlanmasıdır (K. Huang vd., 2020; Sachdev ve Gupta, 2019). Bir hastalık için güvenli ve etkili tedaviler bulmak, büyük ölçüde bu moleküler terapötik hedeflerin doğru bir şekilde tanımlanmasına bağlıdır. (K. Huang vd., 2021). İlaçların terapötik etkileri İHE aracılığıyla gerçekleştiğinden hedefin de ayrı bir önemi vardır. Bir hedef hastalıkla doğrudan ilişki bir protein, gen veya dolaylı olarak

hastalığa neden olan hedefe karşı koymaya yardımcı bir protein de olabilmektedir. (Ezzat vd., 2019).

İHE tahmini konusuna çok benzer bir çalışma alanı olan bileşik protein etkileşimi (BPE) tahmini konusu bulunmaktadır. Bu iki çalışma alanı birbirine çok benzer görülse de bu alanları birbirinden ayıran önemli farklar mevcuttur. İHE genel olarak ilaç konumlandırma çalışmaları ve ilaç yan etkilerinin belirlenmesine yardımcı olmak için çalışılmaktadır. İlaç konumlandırma çalışmaları, ilaçların birincil hedefler ile ilişkisi dışında hangi hedefleri etkilediğini belirlemek amacıyla yapılmaktadır. Ayrıca mevcut ilaçların farklı hastalıklar için kullanılmasını destekleyen çalışmalardır. Bu sebeple İHE çalışmalarında ilaç veri seti, kimyasal bileşik verileri dahil olmadığından dolayı daha azdır ve sadece ilaç verileriyle sınırlıdır (Ashburn ve Thor, 2004; Du vd., 2022). Bu tez çalışmasında da bu bilgiler doğrultusunda hali hazırda mevcut FDA onaylı ilaç veri setiyle çalışılmıştır.

Son yıllarda tüm dünyanın tecrübe ettiği Covid-19 pandemisi sürecinde de hızlıca keşfedilip, geliştirilebilen; etkili, maliyeti düşük ilaç ve aşıların piyasaya sürülmesi çok elzemdir(Xu vd., 2021). Sağlık alanında hesaplamalı yöntemlerin geliştirilmesi ve klasik yöntemlere entegre edilmesiyle ilaç keşfi süreçlerinin daha verimli hale geleceği çok açıktır. Bu doğrultuda, ilaç keşif çalışmalarının geliştirilmesinin hız kazanması için bir hedef ile ilacın etkileşiminin bilinmesi gerekmektedir. Klasik ilaç keşfi yöntemlerinde tek gen-tek protein-tek hastalık görüşüyle ilerleme kaydedilmesi Alzheimer hastalığı gibi karmaşık yapılı rahatsızlıkların tedavi aşamalarında ciddi sorunları beraberinde getirmekle birlikte doğru etkiye sahip olamamalarını da açıklamaktadır (Hopkins, 2008). Ayrıca bu yöntem ile elde edilen ilaçlar çok faktörlü ve çoklu hedefin etkilediği hastalıklara karşı savunmasız kalırlar. Sonuçta çözüm oluşturulamadığı için tüm sürecin yeniden baştan başlaması gerekmektedir ki bu uzun ve pahalı süreç oldukça yorucudur. Bu yüzden ilaç yeniden konumlandırma çalışmaları son yıllarda ilaç sektörünün ve araştırmacıların odak noktası olmaya başlamıştır.

İlaçların yeniden konumlandırılması, mevcut ilaçların yeni endikasyonlar için yeniden kullanılmasıdır; yani mevcut ilaçlar, başlangıçta geliştirildikleri hastalıklardan başka hastalıkları tedavi etmek için kullanılabilir (Ezzat vd., 2018). Yakın zamanda

tanımlanmış olan İHE, halihazırda onaylanmış ilaçlarla etkileşime giren yeni hedeflerin yanı sıra belirli hastalıklarla ilgili genleri hedefleyen yeni ilaçların belirlenmesi için de gereklidir (Ezzat vd., 2018). İHE tahmini alanındaki gelişmeler ilaç yeniden konumlandırılmasının önünü açmaktadır. Çünkü FDA onayı alınmış ilaçların tekrar farklı bir hastalık için yeniden konumlandırılıp kullanılması birçok maliyet ve prosedürü kısaltmakla birlikte hızlı bir tedavi yolunu sağlamaktadır. Bu alanın en güzel örneği Gleevec (imatinib mesylate) ilacıdır. Bu ilaç piyasaya sürüldüğünde, yalnızca lösemiye bağlı Bcr- Abl füzyon geni ile etkileşime girdiği bilinmektedir fakat sonrasında aynı ilaç gastrointestinal stromal tümörleri tedavi etmek için yeniden konumlandırılmıştır. Ayrıca PDGF ve KIT ile etkileşime girdiği gösterilmiştir. Gleevec, ilaç yeniden konumlandırmanın önemini açıkça gösteren ilaçlardan biridir ve başarısıyla da bu çalışmaların devamı için ön ayak olmuştur (Druker, 2002). Bu sebeple, İHE tahmininin önemli katkıları Alzheimer gibi çeşitli hastalıklar için geliştirilmesinin önünü açmıştır.

Demansın sıklıkla karşılaşılan şekli olan Alzheimer hastalığı(AH), beyinde fibröz amiloid beta proteininin birikmesiyle, entelektüel ve sosyal yeteneklerde kayıp gibi klinik semptomların oluşmasına ve beyin hücrelerinin yavaş ölümüne sebep olan nörodejeneratif bir hastalıktır (Lane vd., 2018). Amiloid β (Aβ) formlarının hücre dışı boşluklarında, kan damarı duvarlarında oluşumu ve mikrotübül proteini olan Tau'nun nöronların nörofibriler yumaklarında oluşan birikimi Alzheimer hastalığının belirtileridir. Hastalığın ortalama süresi 8-10 yıldır fakat 20 yıl süren uzun preklinik ve prodromal evreler de görülmektedir (Masters vd., 2015). Dünyada ki mevcut sonuçlara göre AH, 44 milyon kişiyi etkileyen ve kişide sürekli ilerleme gösterme özelliğine sahip bir rahatsızlıktır (Podder vd., 2018). Hastalığın bazı semptomlarını iyileştirebilen FDA onaylı ilaçlar bulunsa da günümüzde hala hastalığın altta yatan mekanizmalarını ve hastalığın seyrini değiştirebilen ilaçlar mevcut değildir (Masters vd., 2015). Bu bağlamda, tez çalışmamızda ilaç yeniden konumlandırma çalışmalarına katkı sağlamak amacıyla İHE tahmini yapılmak üzere hedef veri seti için AH'na ait proteinler seçilmiştir.

Bu tez çalışmasının İHE tahmini alanına katkıları şunlardır;

1. İlaç temsili için ilaçların iki boyutlu moleküler yapı görüntülerinden transformers ağı kullanılarak özniteliklerinin çıkartıldığı, hedef temsilinin protein vektörleri ile oluşturulduğu ve İHE temsili verileri ile birleştirilen bir model olmasıdır.

2. İHE çifti veri setinde bilinmeyen etkileşimlerin, birbirine en uzak protein temsillerinin Öklid, Manhattan, Minkowski uzaklık yöntemleri ile belirlendiği ve ilgili proteine ait ilaçların negatif örnek kabul edilerek etiketlenmesidir. Çünkü literatürde İHE tahmini için kullanılan bilinmeyen etkileşimler sıklıkla rastgele oluşturulmaktadır ve bu durum büyük bir probleme yol açmaktadır. Tez çalışmasında; Alzheimer hastalığının verileri kullanılarak ilaç hedef etkileşimi tahmini gerçekleştirilmiştir.

## 2. KAYNAK ÖZETLERİ

Bu bölümde, araştırmada ele aldığımız ilaç hedef etkileşimi(İHE) tahmini alanında var olan problemlerin çözümleri için çeşitli stratejilerin bulunduğu makaleler ve ilgili yöntemlerin sonuçlarını etkileyen yayınlar incelenecektir.

Pahikkala vd. (2015), çalışmasında bir makine tahmin modeli olan en küçük kareler yöntemini kullanarak ilaç hedef bağlanma afinitesini tahmin etmede oldukça doğru sonuçlar elde edildiğini göstermişlerdir.

Hu vd. (2016), İHE tahminini bir ikili sınıflandırma problemi olarak ele aldıkları çalışmasında, bir proteinin ilgili ilaçla etkileşime girip girmeyeceğinin tahmininin yapıldığı bir model oluşturmuşlardır. Yöntemlerinde, ilaç ve hedeflerin özelliklerini yapılandırmak amacıyla oto-kodlayıcılar ile özniteliklerini çıkarttıkları MFDR adlı modeli önermişlerdir.

He vd. (2017), çalışmasında kimyasal bileşiklerin ve proteinlerin bağlanma yakınlıklarının sürekli verilerini (ikili olmayan) tahmin eden SimBoost yöntemini önermişlerdir.

Wen vd. (2017) , FDA onaylı ilaç ve hedeflerin ham girdi verilerini etkili bir şekilde temsil etmek ve bunun yanında etiketlenen İHE'leri doğru bir şekilde tahmin etmek için oluşturdukları yaklaşımlarında derin inanç ağlarını (DBN) kullanmışlardır.

Ezzat vd. (2017), İHE tahmini için hem girdi verilerinin boyutlarını azalttıkları hem de topluluk öğrenmesi (ensemble learning) teknikleri; karar ağaçları (Decision Tree) ve Kernel Ridge Regresyon kullandıkları çalışmalarında, İHE tahmini alanında yardımcı olabilecek bir araç geliştirdiklerini göstermişlerdir.

Öztürk vd. (2018) , ilaç hedef bağlanma afinitesini ölçmek için, hedef ve ilaç verilerinin bir boyutlu temsillerini evrişimli sinir ağları(CNN) ile modellemişlerdir. İHE tahmini için derin öğrenme tabanlı modellerin kullanıldığı ilk yayınlardan biri olan DeepDTA, KronRLS algoritması ve SimBoost yöntemi ile karşılaştırıldığında çok daha iyi bir performans göstermesiyle de ilgi çekmiştir.

Y. Huang vd. (2018), çalışmasında Yamanishi vd. (2008), altın standart veri setini kullanmışlardır. Hedef verilerini Pseudo-SMR tanımlayıcısı ile temsil etmiş ve ilaçlar için SMILES dizilerini moleküler parmak izlerine dönüştürdükleri bir rastgele orman modeli geliştirmişlerdir.

Lee vd. (2019), ilaç, hedef ve etkileşim verilerinin hepsini ele alan kemogenomik yöntem kullanmışlardır. Bu modelin öne çıkmasının sebebi, protein kalıntılarını yakalamak amacıyla ham protein sekanslarının word2vec algoritmasını temel alan ProtVec vektörlerine dönüştürüldüğü ve sonrasında evrişimli sinir ağlarını (CNN) kullanmaları olmuştur. Çalışmalarında, yeni türetilen protein özellikleriyle birlikte ilaç verilerinin SMILES dizilerini Morgan/Circular parmak izlerine dönüştürülmüş ve sonrasında tam bağlı katmanda birleştirilmesiyle yeni etkileşim ihtimali daha yüksek ilaçlar belirlenmiştir.

Chu vd. (2021), ligandlar ile hedeflerin benzerlik özelliklerini ve bilinen İHE verilerinin kademeli olarak beslendiği derin orman tahmin performansına dayalı İHE-CDF modelini geliştirmişlerdir. Sonucunda AUC, AUPR VE f1 skorlarını değerlendirmişlerdir ve ayrıca ileri çalışmaları gerçekleştirerek DrugBank ve Kegg veri tabanları tarafından onaylanmış 1352 yeni ilacı tahmin etmişlerdir.

Rifaioglu vd. (2020), girdi verileri için ilaçların SMILES dizilerinin moleküler yapılarını Python Rdkit kütüphanesi ile oluşturulmuş ve ilgili kodlar aracılığıyla 200 * 200 piksellik iki boyutlu görüntü gösterimlerini elde etmişlerdir. Oluşturulan iki boyutlu görüntü temsilleri, yeni tahmin edilecek etkileşimlerin aktif ve aktif olmama sonuçlarını üretmek için derin evrişimli sinir ağları ile modellenmiştir. Araştırmanın sonucunda DrugBank veri tabanının öngördüğü yeni İHE'leri tahmin edilmiştir.

K. Huang vd. (2021), ilaçların SMILES bilgilerini ve proteinlerin aminoasit sekans dizi çiftlerini girdi olarak kullandıkları kapsamlı bir kütüphane gerçekleştirerek ön plana çıkmışlardır. İlaçlar için; SMILES dizi bilgilerini, Morgan/Circular Parmakizi, CNN, RNN, transformers, moleküler yapıları için grafik nöral ağlarını, proteinler için ise; AAC, CNN, RNN, transformers ağını ve yapısal parmakizlerini içeren diğer kütüphaneler ile karşılaştırıldığında kolay kullanımı ve yüksek tahmin gücü olan güçlü bir potansitele sahip DeepPurpose adlı derin öğrenme kütüphanesini oluşturmuşlardır.

İlaç hedef bağlanma afinitesini tahmin etmek için derin öğrenme yöntemlerini kullandıkları bu çalışma da Lin vd. (2020), yerel kimyasal bağlamla birlikte moleküler yapıları da dikkate alma amacıyla protein temsili için ProtVec ile Smi2Vec embedding tekniklerini ve ilaçların temsili için de ham SMILES dizilerini kullanmışlardır.

İHE tahmini için Thafar vd. (2020), grafik gömme, grafik madenciliği, benzerlik ve özellik tabanlı teknikleri ile makine öğrenimini birleştirdikleri ilaç-ilaç ve hedef-hedef benzerliklerini de yapıya dahil ettikleri DTIgems+ yöntemini önermişlerdir.

Mahmud vd. (2020), potansiyel İHE'lerini tahmin etmek amacıyla özellik çıkarımı, veri dengeleme, özellik seçilimi ve sınıflandırmadan oluştuğu derin öğrenme tabanlı DeepAction yöntemini önermişlerdir. Yöntemlerinde verideki sınıf dengesizliğini önlemek amacıyla MMIB tekniği ve LASSO modelini kullanmışlardır.

K. Huang vd. (2020), ilaç, hedef ve etkileşim verilerini kullandıkları MolTrans adlı yeni bir yöntem önermişlerdir. İHE tahmini için kullanılacak verilerin moleküler temsillerinden elde edilen özniteliklerinde genellikle çalışmalarda tahmin için işe yaramadığı düşünülen kısımlarını bulunmaktadır. MolTrans'ın ayrıcalığı ise transformers ağı ile eğitilerek bu verilerin işlendiği bir model olmasıdır.

Thafar vd. (2021) , Yamanishi vd. (2008) ait verisetini kullandıkları çalışmasında, İHE tahmini için heterojen ağın kullanıldığı, ilaç ve hedefler için otomatik olarak özellik çıkarımı yaptıkları benzerlik tabanlı DTI2Vec adlı aracı geliştirmişlerdir.

Protein -ligand bağlanma afinitesi tahmini için Wei ve Gong, (2021) ilaçların SMILES dizilerini üzere önceden eğitilmiş gömme (embedding) tekniklerini kullanan Mol2Vec ve ProSE yöntemleriyle yoğun vektör temsillerine dönüştürmüşlerdir. Aminoasit dizilerinin ise ResNet tabanlı bir boyutlu CNN'e beslendiği ve sonrasında biLSTM ile birleştirdikleri bir model önermişlerdir.

G. Liu vd. (2021), ilaçlar için Mol2Vec vektör temsilini, proteinler için ProtVec vektör temsilini, ilaç hedef bağlanma bölgeleri için Bionoi vektör temsilini kullanmışlardır. Araştırmacılar, ek olarak protein-protein etkileşimi ağlarını ve gen ekspresyonu bilgilerini de kullandıkları yüksek veri kapasiteli GraphDTI yöntemini önermişlerdir.

Nguyen vd., (2021), İHE tahmini için oluşturdukları protein temsilleri ile ilaçların SMILES dizilerinden ürettikleri moleküler grafikleri ve ayrıca ilaçları 4 farklı temsile çevirdikleri bir model önermişlerdir. İHE tahmini için farklı yöntemler uyguladıkları modellerinde, diğer çalışmalar ile karşılaştırıldığında daha iyi sonuçlar elde etmişlerdir.

İHE tahmini için benzerlik tabanlı bir model öneren Song vd. (2022), Tanimoto katsayısı, Levenshtein mesafesi, evrişimsel sinir ağı ve transformatör ağını ilaç ve hedeflerin verilerinin alt yapı özellik çıkarımı için kullanmışlardır.

Zhang vd. (2022), ilaç moleküler yapı bilgilerinin yakalanmaması problemini ele aldıkları çalışmasında, hedef dizi bilgilerinde ki yerel kalıntıların belirlenmesinin de önünü açan evrişimli sinir ağlarını kullandıkları bir model geliştirmişlerdir. Bu model diğer yöntemlere göre daha basit ve kullanılabilir bir model olmasının yanı sıra ilaç moleküler yapılarının öznitelik bilgilerini çıkarttıktan sonra grafik nöral ağı ve transformers ağına beslemeleriyle de dikkat çeken farklı bir çalışma olmuştur. Sonuçların ilaç yeniden konumlandırma için kullanılabilirliği ve doğru etkileşim sonuçlarının tahmin edilebilirliği bu çalışmayı yine ön plana çıkaran etmenlerden biridir.

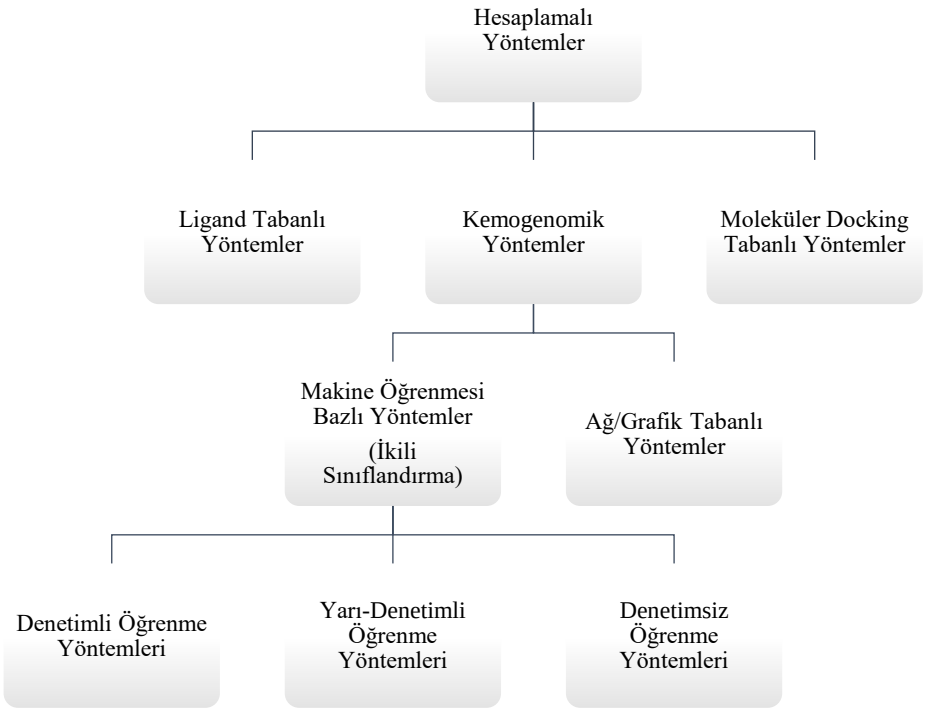
Li vd. (2023), ilaç ve hedef verilerinin yapı özelliklerinden öğrenildiği, daha önce bu alanda kullanılmayan 3 farklı (intrinsic embedding, relational embedding, and annotation embedding) gömme tekniğini kullanarak temsillerini oluşturdukları derin öğrenme tabanlı bir model önermişlerdir. Araştırmaları sonucunda SARS-CoV-2 viral proteinleri için 45.603 adet yeni ilaç- hedef etkileşimi belirlemişler ve tahminleri desteklemek için iki yönlü istatistiksel testler ile de doğrulamışlardır.

# 3. MATERYAL VE YÖNTEM

Bu başlıkta; tez çalışması ile ilgili temel bilgiler, kullanılan veri tabanları, kullanılan Python kütüphaneleri, yöntemler ve çalışmanın genel akışına ait bilgilere yer verilmiştir. Ayrıca, İHE tahmini için kullanılan yöntemler ve Alzheimer hastalığına ait bilgiler incelenmiştir.

## 3.1. İlaç Hedef Etkileşimi (İHE) Tahmini Ve Kullanılan Yöntemler

İlaç keşfinde, yüksek verimli tarama teknolojilerindeki gelişmeler sayesinde artık binlerce bileşiğin aynı anda taranması mümkündür. Fakat protein-hedef kombinasyonlarının çokluğu nedeniyle, hedef ve kimyasal bileşik alanlarının tamamı hala kapsamlı bir şekilde analiz edilememektedir. Bu sorunun çözümü için İHE' lerin in siliko yöntemler ile araştırılması ve süreci hızlandırması gerekmektedir (Atas ve Doğan, 2022). İHE tahmini için geliştirilen hesaplamalı teknikler özellik ve benzerlik tabanlı olarak kabul edilebilir fakat makine öğrenmesi teknikleriyle geliştirilen İHE tahmini için daha etkili hesaplamalı yöntemler mevcuttur. Bu doğrultuda tez çalışmasında makine öğrenmesi bazlı (ikili sınıflandırma) yöntemler kullanılmıştır.



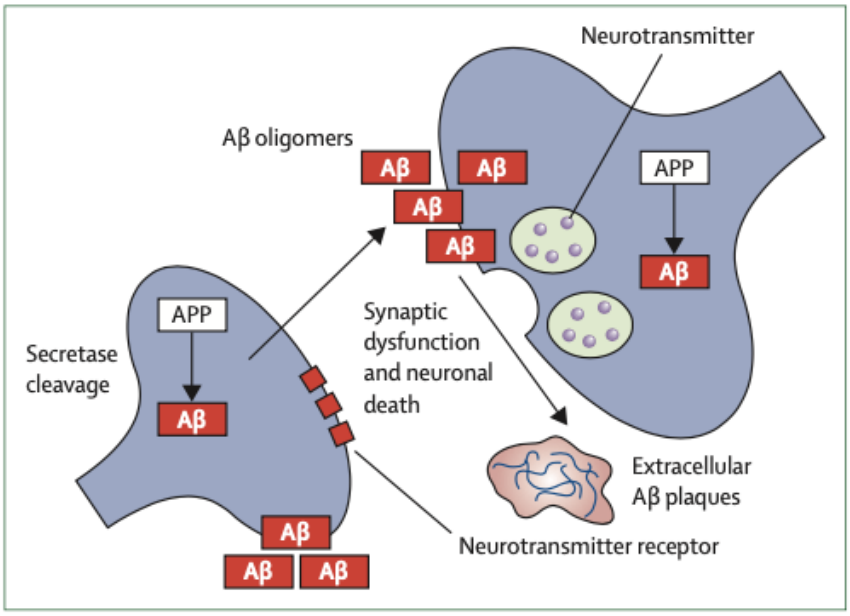

Şekil 3.1. İHE tahmininde kullanılan hesaplamalı yöntemler (Chu vd., 2021).

Hesaplamalı yaklaşımlar, DrugBank, Kegg, CheMBL, BindingDB, UniProt, TTD gibi bilinen İHE verileri içeren büyük çevrimiçi veri tabanlarının var olmasıyla birlikte artık daha yaygın kullanılmaktadır. Büyük ve çeşitli veri tabanlarını kullanan farklı hesaplama algoritmaları, İHE tahmin etme girişimlerini ve mevcut ilaçlar için yeniden konumlandırma/kullanımı çalışmalarını desteklemektedir (Ezzat vd., 2017). Desteklenen İHE tahmini yöntemleri Şekil 3.1.' de gösterilmiştir (Chu vd., 2021).

İHE tahmini yöntemleri; ligand bazlı, moleküler kenetleme(docking) tabanlı ve kemogenomik bazlı olarak 3 ana grupta incelenebilir. Ligand tabanlı metotlar ilaçların benzerliğine dayanmaktadır. Bu metotta, benzer özelliklere ve eğilimlere sahip moleküllerin proteinleri benzer şekillerde bağlanır fikri esastır. Bu yöntemin tercih sebebi, ilacın ilgili proteinin aktif bölgelerine bağlanmasıdır. Fakat bu çalışma proteine ait etkileşimi olan ilaç bulunmadığında kullanılamamaktadır. Moleküler docking tabanlı yöntemlerde, ilaç ve proteinin üç boyutlu yapısı alınır, ardından etkileşime girip giremeyeceklerini belirlemek üzere simülasyon programı üzerinde çalışılır. Bu yöntemin avantajı, hedef proteinler için çok zengin bilgi ile çalışılması ve daha gerçek sonuçların elde edilmesidir. Fakat her proteinin üç boyutlu yapısının bulunmaması bir dezavantajdır ve üç boyutlu yapısı olmayan proteinlerin yapı tahminin araştırılma süreci zorludur. Kemogenomik yöntemlerde ise, ilaçlara ve hedefe ait verilerin eş zamanlı kullanıldığı İHE tahmini yapılmaktadır. Bu metodun en önemli avantajı, bol miktarda ve kolayca erişilebilen veri ile çalışılmasıdır. Ek olarak kemogenomik yöntemlerin başarısı diğer yöntemlere göre çok yüksek olduğu için tercih sebebidir. Bu yöntemin dezavantajı ise bilinen İHE ve doğrulanmamış negatif İHE örneklerinin azlığıdır (Chu vd., 2021; Ezzat vd., 2018; Kaushik vd., 2020; W. Zhou vd., 2016).

### 3.2. Alzheimer Hastalığı(AH) ve Amiloid Precursor Proteini(APP)

Alzheimer hastalığı ilk olarak 1907 yılında Alois Alzheimer'ın raporladığı Auguste Deter adlı hastanın otopsi sonuçlarına göre bildirilmiştir. Raporda belirtildiği ve sonra ki yıllarda da doğrulandığı üzere nörofibriller yumaklar (NFT) ve amiloid plaklar, Alzheimer patogenezine ait iki temel tanımlayıcıdır (Lane vd., 2018; O'Brien ve Wong, 2011). Tanımlayıcıların yanı sıra temel hastalık süreçlerinin yaşla ilgili, hayat tarzı, çevresel faktörler ile ilgili koruyucu ve hastalığı teşvik edici değişkenlerle etkileşime girmesi muhtemeldir.



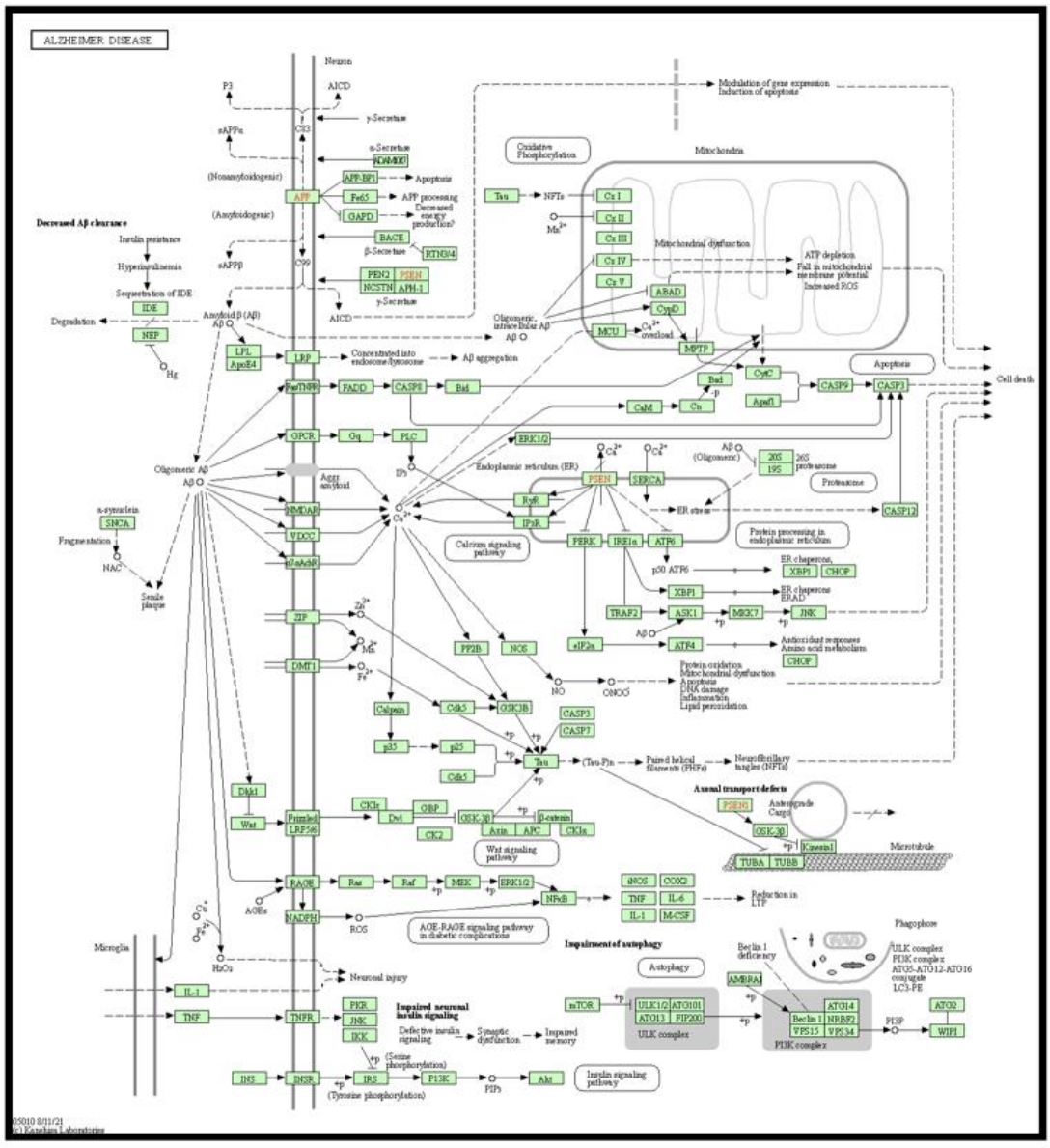

Şekil 3.2. Amiloid kaskad hipotezi (Ballard vd., 2011).

Şekil 3.2.'de amiloid kaskad hipotezine ait bir görsel bulunmaktadır (Ballard vd., 2011). Bu hipoteze göre, amiloid precursor proteini(APP) plakları oluşturmadan önce nöronal hücrelerin hem içinde hem de dışında oluşan amiloid β (Aβ) 'ya dönüştürülür. Amiloid kaskad teorisine göre, dönüştürülen Aβ birikintileri toksiktir bu sebeple işlevsiz sinapslara ve nöronal hücrelerin ölümüne yol açar. Fakat son yıllarda araştırmacılar, amiloid β (Aβ) ve tau birleşik varlığıyla tanımlanan Alzheimer hastalığı ile ilgili orijinal amiloid hipotezinde öne sürülen basit doğrusal nedensellik varsayımından yavaş yavaş uzaklaşmaktadırlar (Ballard vd., 2011; Scheltens vd., 2016). Bu sebeple yeni yaklaşım ve tedavi yöntemlerine ihtiyaç bulunmaktadır.

Şekil 3.3.'de Alzheimer hastalığı moleküler yolaklarına ait görsel bulunmaktadır (Kanehisa, 2002). Alzheimer hastalığı için, amiloid beta plakları ve tau proteinlerinin keşfinden sonra bilim dünyasında çeşitli moleküler yolaklar hakkında bilgi sağlansa da bugün hala hastalık ile ilgili bilgiler kısıtlıdır. Alzheimer hastalığını ve hastalığın patofizyolojisini değiştiren veya hastalığı gerileten tedaviler henüz erişilebilir olmamakla birlikte hastalığın tedavisinde aşılması gereken en büyük sorundur (Lane vd., 2018).

Alzheimer hastalığına ait kesin bilgilerin açığa kolayca çıkabilmesinde protein-protein etkileşimlerine ve ilaç-hedef etkileşimlerine dolayısıyla hesaplamalı yöntemlere

ihtiyaç bulunmaktadır. Bu bilgiler doğrultusunda tez kapsamında, Alzheimer hastalığı verileriyle İHE tahmini alanında ilerlenmiştir.




Şekil 3.3. Alzheimer hastalığına ait moleküler yolaklar (Kanehisa, 2002).

### 3.3. Veri Seti

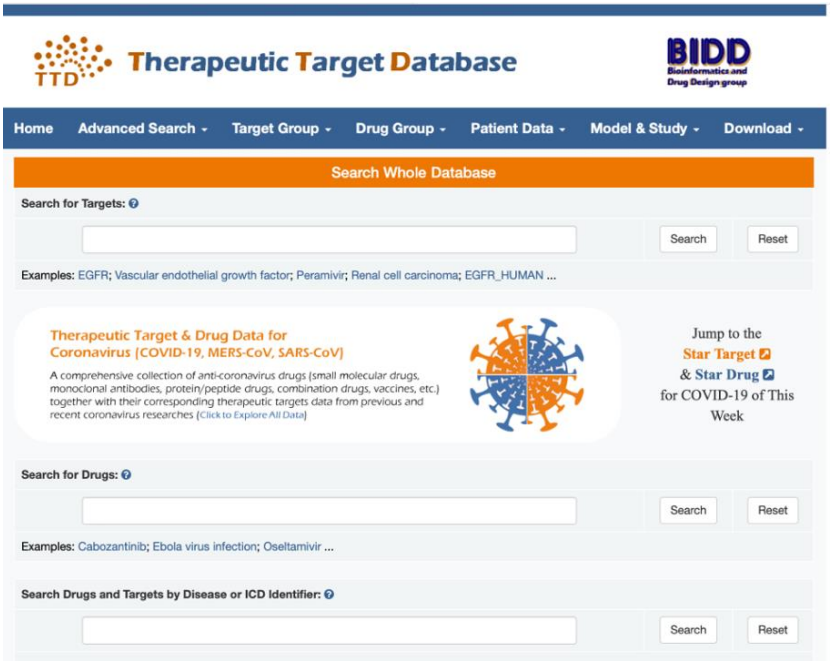
Terapötik hedef veri tabanından (TTD), Alzheimer hastalığı ile ilişkilendirilmiş APP proteini dahil 135 hedef protein seçilmiş ve DrugBank veri tabanından FDA onaylı hali hazırda kullanılan 2465 ilaç verisi elde edilmiştir. İHE verisi için hedef proteinlerin etkileşimi bilinen ilaçlar yine DrugBank veri tabanından elde edilmiştir. Bilinen etkileşimler 1 ile, bilinmeyen etkileşimler ise 0 ile etiketlenmiş olup bu

verilerin rastgele seçiminin yanında; Öklid, Manhattan ve Minkowski uzaklık yöntemleri ile belirlenmiş ilaçlar seçilmiştir. Seçilen ilaç, protein, ilaç hedef etkileşimi verilerinin temsilleri ve özellik çıkarımları ile ilgili literatürde kapsamlı bir şekilde tarama yapılmıştır. Literatürde çok çeşitli yöntemlerin uygulandığı görülmüştür. Bu sebeple, yapılan çalışma için elde edilen veriye en uygun yöntemler belirlenmiş, Python programa dili ve kütüphaneleri ile çalışılmıştır. Kullanılan yöntemlere ait ayrıntılar aşağıda belirtilmiştir.

### 3.4. Kullanılan Veri Tabanları

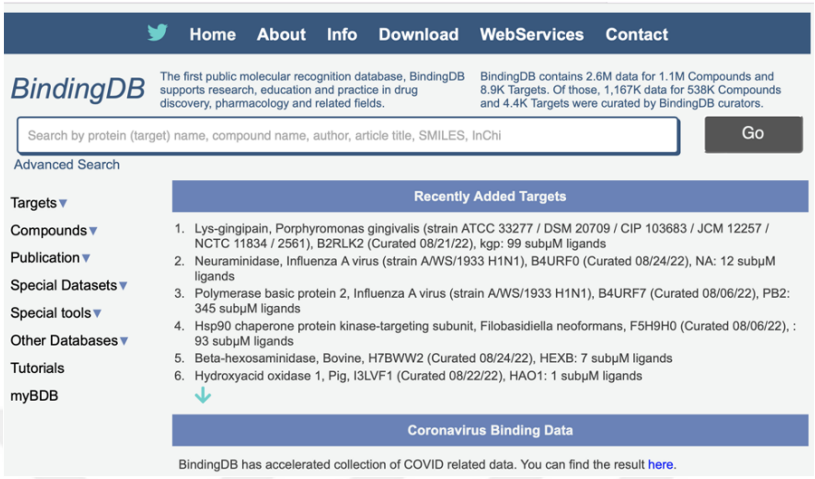
İHE tahmininde bugüne kadar çevrimiçi ve arayüzü kullanıcılara açık bir çok veri tabanı bulunmaktadır. Çalışmanın ihtiyacına göre hangi verilerin gerekli olduğu ve hangi veri tabanının kullanılacağı belirlenir.

Terapötik hedef veri tabanı (TTD) (https://db.idrblab.net/ttd/), literatürde tanımlanan terapötik hedeflerin ve kombinasyonlarının verilerini, gen ifadelerini, hedeflerin hastalık bilgilerini, yolak bilgileri ve ilaç dirençlerinin ayrıntılı bilgilerini içerir. Şekil 3.4. ‘de TTD veri tabanına ait görsel verilmiştir (Y. Zhou vd., 2022). Alzheimer hastalığı ile ilgili hedef proteinler TTD veri tabanından elde edilmiştir.



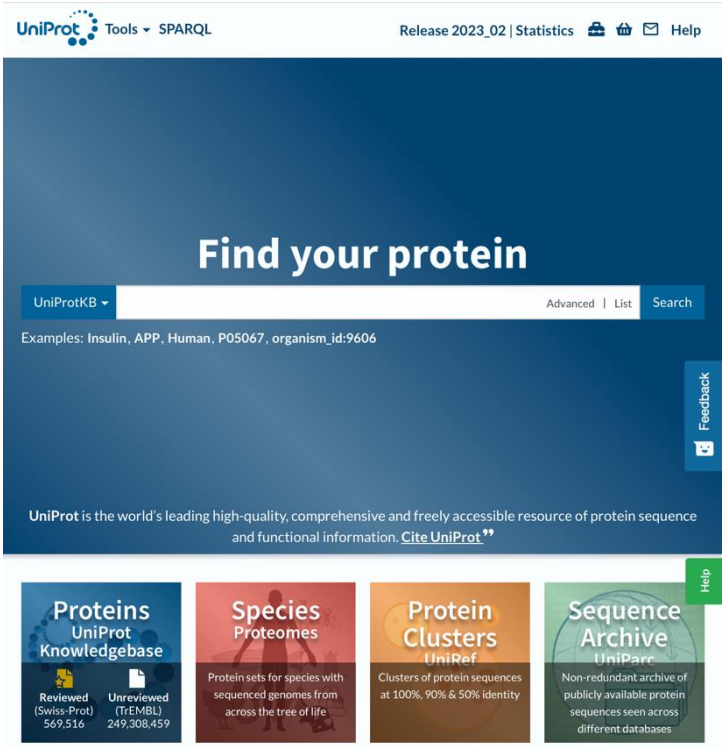

Şekil 3.4. Terapötik target veri tabanının gösterimi (Y. Zhou vd., 2022)

BindingDB (https://www.bindingdb.org/rwd/bind/index.jsp) veri tabanının öne çıktığı alan, hedef ve protein bilgilerinin yanında bağlanma afinite değerlerini içermesidir. Şekil 3.5'de veri tabanına ait görsel belirtilmiştir (T. Liu vd., 2007).



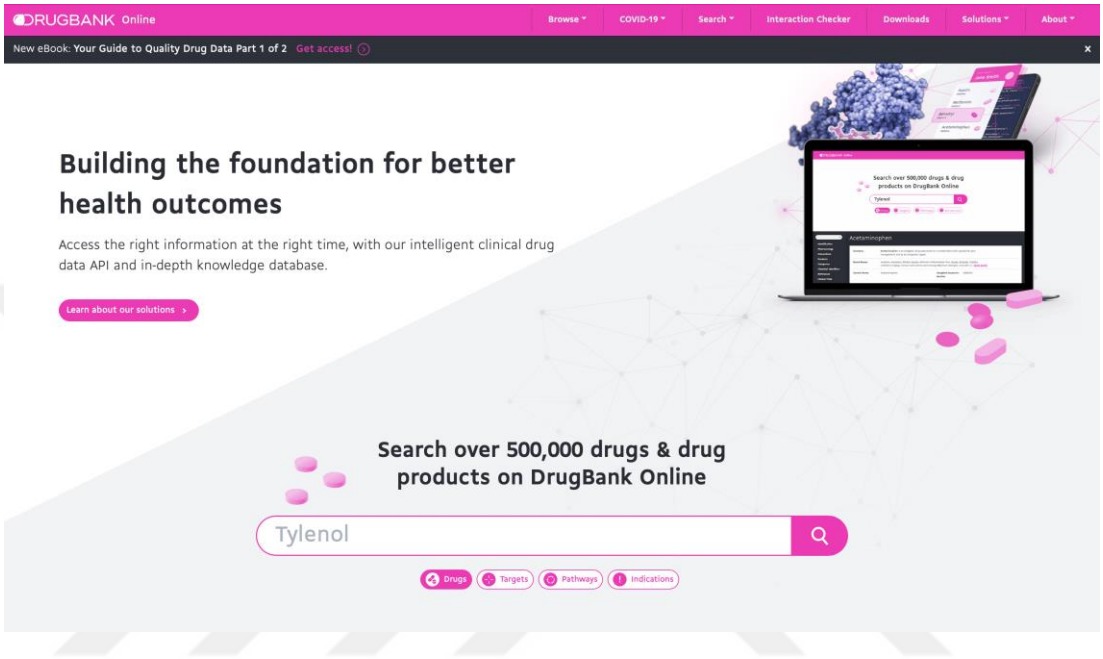

Şekil 3.5. BindingDB veri tabanının gösterimi (T. Liu vd., 2007)

UniProt (https://www.uniprot.org) veri tabanı, 120 milyondan fazla proteine ait bilgileri ve sekanslarını içeren çevrimiçi bir kaynaktır. Proteinlere ait tüm sekans bilgileri UniProt veri tabanından elde edilmiştir. Şekil 3.6. 'da veri tabanı görseli belirtilmiştir (Consortium, 2019).




Şekil 3.6. Uniprot veri tabanının gösterimi (Consortium, 2019)

DrugBank(https://go.drugbank.com) veri tabanı, kapsamlı moleküler bilgiye sahip, onaylı ve deneysel tüm ilaçları, hedefleri, onların tüm etkileşimlerini, mekanizmalarını içeren çevrimiçi çok ayrıntılı bilgiye sahip bir veri tabanıdır. İlaçlar ve etkileşimlere ait bütün bilgiler DrugBank veri tabanından elde edilmiştir. Şekil 3.7. 'de DrugBank veri tabanına ait web sitesi belirtilmiştir (Wishart vd., 2018).




Şekil 3.7. DrugBank veri tabanının gösterimi (Wishart vd., 2018)

### 3.5. Protein Vektörleri (ProtVec)

Doğada yaygın olarak bulunan 20 aminoasidin primer(birincil), sekonder(ikincil), tersiyer(üçüncül) ve kuaterner(dördüncül) yapıları mevcuttur. Proteinlerin birincil yapılarının verilerinin kullanıldığı çalışmalar olduğu kadar üç boyutlu yapı bilgisi bulunan proteinler ile çalışmalar da mevcuttur. Proteinlerin üç boyutlu yapılarının kullanıldığı yöntemler oldukça doğru sonuçlar oluştursa da her proteinin üç boyutlu yapı bilgisinin elde edilebilmesi için fazla bütçeye ve daha fazla zamana ihtiyaç vardır ki bu sebeple her proteinin üç boyutlu yapısı bilinmemektedir ve her proteine ait bu verilerin olmayışı bir dezavantajdır. Bu yüzden son yıllarda uygulanabilirliği artan yöntemler olan proteinlerin aminoasit dizi bilgileriyle oluşturulan temsillerin kullanımı daha yaygındır (Atas ve Doğan, 2022). Çalışmamızda, bu bilgiler doğrultusunda hedef verisi için, proteinlerin birincil yapıları olan aminoasit dizileri ile çalışılmıştır. Her protein için aminoasit dizilerinin farklılığı onları eşsiz ve özel kılsa

da her birinin uzunlukları da aynı zamanda birbirinden farklıdır. Örneğin bir aminoasite ait sekans dizi MSIIGATRLQNDKSD…….. şeklinde başlayabilir ve bir uzunluğa sahiptir, bunun gibi farklı kombinasyonlar ile oluşabilecek proteinlerin aminoasit dizileri de 10 ile 10.000 aminoasit uzunluğunda olabilmektedir. Yapılan çalışmaların çoğunluğunda 2048 amino asit dizisinden kısa dizilerden meydana gelen hedefler mevcuttur ve bu yönden de çalışmamızın avantajlarından biri protein temsilinin 2048 amino asitlik uzun diziler ile de çalışılabilirliğidir.

Biyoinformatik alanında ise sıkça kullanılan aminoasit dizilerinin gösterimleri bilgisayar tabanlı uygulamalarda bilgisayara tanıtılacak düzeyde sayısal verilerden oluşmamaktadır. Bu sebeple aminoasit dizilerinin kullanıldığı, İHE tahmini gibi alanlarda kullanılacak olan protein temsilleri için sayısal vektörlere dönüştürülen yöntemlere ihtiyaç duyulmaktadır. Protein vektörleri, gen vektörleri gibi biyolojik sekansların gösterimlerini betimlemek için oluşturulan biyolojik vektörler biyoinformatik alanında bu açığı oldukça kapatmaktadır. Özellikle protein vektörleri; etkileşim tahmini, yapı tahmini, protein aile sınıflandırılması, protein görselleştirilmesi gibi alanlar için sıklıkla kullanılmaktadır. Bu sebeple tez çalışmasında hedef verisi için, protein aminoasit dizilerinden oluşturulan protein vektörleri (ProtVec) kullanılmıştır. ProtVec yaklaşımında sekanslar, biyokimyasal ve biyofiziksel özelliklerinin karakterize edilmesi amacıyla sinir ağlarını kullanarak n-boyutlu vektörlere dönüştürülür. Bu yaklaşımın temel mantığı doğal dil işlemeye (NLP) dayanır. Benzer kelimeler yakın vektörlere sahiptir. Bu tür vektörleri eğitmenin temel ilkesi, bir kelimenin anlamının bağlamı veya etrafındaki kelimeler tarafından karakterize edilmesidir. Sonuç olarak, kelimeler ve bağlamları pozitif eğitim örnekleri olarak kabul edilir. Skip-gram modeli gibi sinir ağı mimarileri dahil olmak üzere çok sayıda teknik, büyük hacimlerde metin girdisi kullanarak bu tür vektörleri eğitmek için kullanılabilir (Asgari ve Mofrad,, 2015). Bu yaklaşım, aminoasitlerin 3-mer'li ayrılması sürecinde de geçerlidir ve sonucunda n-boyutlu vektörler oluşturulur.

### 3.6. SMILES Dizileri

İlaçların en önemli özelliklerinin belirlendiği kimyasal yapılarının temsilleri bir boyutlu, iki boyutlu ve üç boyutlu olarak tanımlanabilir(Du vd., 2022). Bu tanımlamalardan öncelikle ilaçların iki boyutlu kimyasal yapılarının elde edilmesi için

bir boyutlu temsilleri olan SMILES dizileri kullanılmıştır. Bu gösterim ilk olarak Weininger (1988) tarafından bileşik ve ilaçların iki boyutlu kimyasal yapı formlarının bilgisayarlar tarafından daha kolay ve anlaşılabilir olması amacıyla SMILES(Simplified-Molecular-Input-Line-Entry-System) adı verilen dizilere dönüştürüldüğü kimyasal bilgi sistemleridir. SMILES dizileri; ilaçlar ve bileşikler de ki atomlardan en karmaşık yapılara kadar çevrilebilen, kimyasal yapılarının grafik görüntülerini basitleştirilmiş şekilde sunabilen ve belirli kurallara sahip son yıllarda da ilaç keşfi, İHE tahmini gibi alanlarda ilaç temsili için sıklıkla kullanılan gösterimlerdendir (Weininger, 1988). İlaçların SMILES dizilerinin moleküler yapılarına dönüştürülmeye ihtiyacı vardır. Bu sebeple Python yazılım programı kütüphanelerinden biri olan RDKit kütüphanesi kullanılmıştır.

### 3.7. RDKit Kütüphanesi

C++ ve Python tabanlı RDKit kütüphanesi, kemoinformatik ve biyoinformatik alanında çalışan araştırmacılar için oluşturulmuştur. Tez çalışmasında ilaç temsili için ilaçların SMILES verilerinden moleküler yapılarının oluşturulması amacıyla RDKit kütüphanesi kullanılmıştır. RDKit kütüphanesi, keminformatik alanında çalışan araştırmacılara çok çeşitli işlevler sunmaktadır. Kimyanın informatik bilgilerini içeren, görselleştirmede yaygın kullanılan, SMILES, SDF VE PDB gibi dosya formatlarıyla kolay çalışılabilirliği, ilaç keşfi ve tasarımı için uygunluğu açısından kullanışlı bir kütüphanedir.

### 3.8. Evrişimsel Sinir Ağları (CNN)

Konvolüsyonel Sinir Ağları (CNN), canlıların çevrelerini doğal olarak nasıl algıladıklarından ilham alan, 1980 yılında keşfedilen ve iyi bilinen bir derin öğrenme mimarisidir. Araştırmacılar, derin öğrenmenin gücünden yararlanarak kimyasal yapılar ve protein dizileri içindeki karmaşık kalıpları ve korelasyonları yakalamak için CNN'leri kullanabilirler. CNN, ilaçların hedefleriyle nasıl etkileşime gireceğini tahmin etmeyi mümkün kılarak olası ilaç adaylarını bulmayı, moleküllerin nasıl çalıştığını anlamayı ve ilaç keşif sürecini hızlandırmayı kolaylaştırır (Gu vd., 2018).

### 3.9. Transformers Ağı

Transformer ağı, doğal dil işleme (NLP) ve görüntü tanıma gibi alanlarda son yıllarda sıkça kullanılan derin öğrenme modellerinden biridir. İlaçların moleküler yapılarından oluşan iki boyutlu görüntüleri farklı boyutlardaki pikselleri içerir ve her pikselin değeri ilacın bir özelliğini ifade eder. İfade edilen özellikler atomik bağlardan, fonksiyonel gruplardan ya da diğer kimyasal özelliklerden oluşabilir (Zhang vd., 2022).

İlaç temsili ve ilaç keşfinde de başarılı sonuçlar veren transformers ağı, ilaç molekülünün moleküler yapısındaki çeşitli atomlar arasındaki karşılıklı bilgiyi daha derin bir düzeyde çıkarmak için kodlayıcı(encoder) ve çözücü(decoder) içeren bir dizi dikkat mekanizmasını kullanmaktadır. Her piksel ile komşuları arasındaki bağlantıları ve korelasyonları dikkat mekanizması tarafından analiz edilir. Moleküler yapıdaki karmaşık etkileşimleri ve kalıpları analiz etmek için, pikseller arasındaki çok uzak bağlantıları ve bağımlılıkları modellenir.

### 3.10. İlaç- Hedef Etkileşimi Verisi

İlaç hedef etkileşimi(İHE) verilerinin elde edilmesi için DrugBank veri tabanı kullanılmıştır. Her hedefe ait farmakolojik olarak pozitif etkileşim gösteren ilaç veri bilgileri ilgili veri tabanından indirilmiş ve 1 ile etiketlenmiştir. Negatif etkileşim verisine ait ilaçların rastgele seçilmesinin yanında en önemli uzaklık ölçüm metodlarından olan Öklid, Manhattan ve Minkowski yöntemleri ile belirlenen ilaçlar 0 ile etiketlenmiştir. Böylece etkileşim veri seti de tamamlanmış olmaktadır.

İHE tahmini alanında, bilinmeyen İHE verileri negatif etkileşim verisi olarak kabul edilebilir ve bu alanın çalışmasını sıkıntıya sokan en büyük sorunların başında gelmektedir. Negatif etkileşim verilerini seçmenin zor olmasının sebebi doğrulanmış İHE dışı verilerin belirli olmamasıdır. Genellikle negatif veriler bilinen İHE etkileşimi çiftlerinin haricinde kalan alandan rastgele seçilmektedir. Rastgele seçilen veriler arasından negatif etkileşim veri setinin hazırlanmasının temelinde yatan problem ise kabul edilen negatif etkileşim verisinin yıllar sonra pozitif etkileşim verisi olarak keşfedilmesinin de mümkün olabileceği ihtimalidir. Böyle bir durumda negatif/bilinmeyen etkileşimlerin bilinmemesi çok büyük bir eksiklik olmakla birlikte

belirlenmesi de çok mühimdir (Peng vd., 2020). Bu bağlamda, konu üzerine çeşitli yöntemler geliştirilmiş ve hala geliştirilmeye bu alanın zorluklarının kısıtlanması üzerine çalışılmaktadır.

Öklid uzaklık yöntemi; makine öğrenmesi, kümeleme, sınıflandırma, veri bilimi, veri madenciliği vd. alanlarda sıkça kullanılan, iki vektör, nokta, görüntü vb. araçların arasında ki en kısa mesafenin ölçümünün yapıldığı ve genelde bu noktaların arasındaki farklılıkların ve benzerliklerin hesaplamasında kullanılan bir yöntemdir (Danielsson, 1980; Taşcı ve Onan, 2016).

Manhattan uzaklık methodu, Öklid uzaklığına benzer olan ve ortak uygulama alanının bulunduğu bu yöntem de iki vektör, nokta, görüntü, piksel vb. araçların arasındaki en uzak mesafenin ölçümünün yapıldığı hesaplama yöntemidir (Danielsson, 1980; Taşcı ve Onan, 2016).

Minkowski uzaklık yöntemi ise, bir değişkene bağlı olarak uzaklık mesafesinin ölçüldüğü ve önceki iki uzaklık metodunun genelleştirilen halidir (Singh vd., 2013; Taşcı ve Onan, 2016).

### 3.11. Kullanılan Makine Öğrenmesi Algoritmaları

İlaçlar için CNN ve transformers ağı kullanılarak oluşturulan öznitelikler ile hedefler için oluşturulan protein vektörleri; İHE tahmini için rastgele orman, Destek Vektör Makineleri (DVM) vb. gibi çeşitli makine öğrenme modellerine beslenir. Tahmine dayalı modeller genom, transkriptom, proteom gibi biyoinformatik analizlerde ,tıbbi teşhisin konulmasında, prognostik gibi görevlerde sıklıkla kullanılmaktadır. Tahmine dayalı modellerin kullanıldığı çalışmalarda ise hipotezlerin doğruluğunun değerlendirilmesi gerekmektedir. Bu bağlamda; rastgele orman(RO), lojistik regresyon(LR), karar ağaçları(KA), destek vektör makineleri(DVM) gibi modelleme süreci gerçekleşir ve tahmine dayalı modellerin gerçekleştirildiği araştırmaların sonuçları değerlendirilir.

İHE tahmininde tez kapsamında kullanılan makine öğrenme algoritmaları doğrultusunda çalışılan alan bir ikili sınıflandırma problemi (binary classification)

olarak ele alınmıştır. İHE tahmin edilmesi için, kemogenomik yöntemler eşliğinde bir makine öğrenimi sorunu olarak aşağıda belirtilen sınıflandırıcılarla modellenmiştir.

### 3.12. Rastgele Orman

Rastgele orman (RO), birçok karar ağacının entegre edilmesiyle oluşturulan bir sınıflandırıcıdır. RO, Bireysel ağaç tahminlerinin fikir birliğine dayanarak nihai tahmin yapılır. Güçlü bir sınıflandırma yeteneğine sahip olan RO; tahmin, değişkenlerin önemi, boyutsallık azaltmasında ve anormal nokta tespiti gibi çeşitli problemlerde sıkça kullanılmaktadır. Özellik sayısının örnek sayısından fazla olduğu durumlarda aşırı uyum sorunu oluşturmaması ve yüksek boyutlu veri kümeleri üzerinde iyi performans ve doğruluk sağlaması sebebiyle güçlü avantajlara sahiptir. (Sachdev ve Gupta, 2019; Shi vd., 2019). Bu sebeple, İHE tahmini araştırmaların da ikili sınıflandırma görevlerinde RO sıklıkla kullanılmaktadır.

### 3.13. Lojistik Regresyon

Bir çok veri sınıflandırma görevinde tercih edilen modellerden biri olan Lojistik regresyon (LR), yanıt değişkenlerini ikili (binary) bir sınıf olarak tahmin etmek için diğer değişkenlerdeki gözlem değerlerini kullanmaktadır. Değişkenlerin ikili sınıflandırma olarak kullanılması, İHE tahmini etkileşim verileri açısından bir avantaj sağlar. LR sınıflandırıcı modeli de diğer modeller gibi İHE tahmininde sıklıkla kullanılmaktadır. (Dreiseitl ve Ohno-Machado, 2002).

### 3.14. Karar Ağaçları

Karar Ağaçları (KA), verilerin ayrılmasını optimize eden bir kritere göre periyodik olarak bölünür ve sonucunda ağaç benzeri bir yapı ortaya çıkmaktadır. Karar ağaçlarının RO'dan farkı, giriş örneğinin özellik değerlerine dayalı olarak, bir sonucu tahmin etmek için ağaç yapısını kökten bir yaprak düğüme doğru dolaşır. Yaprak düğümün sınıf etiketi tahmini için temel teşkil eder. Sonuç olarak, KA ve RO modelleri sınıflandırma (çoğunluk oylaması) ile regresyon(ortalamalar) görevlerinde sıklıkla kullanılan araçlardır. (Dreiseitl ve Ohno-Machado, 2002).

### 3.15. Destek Vektör Makineleri

Destek Vektör Makineleri (DVM), veri analizi sürecini kolaylaştırmak için verileri eğitim örneklerine göre kategorilere ayırırlar. DVM, çeşitli veri madenciliği uygulamalarında kullanılan ve yöntemin kolay uygulanması ve karmaşıklığının az olmasından dolayı tercih sebebidir. DVM'ler, İHE tahmininde ilaç hedef çiftlerini etkileşime giren ve girmeyen gruplara ayırmanın bir yolunu sunduğu ve ayrı ayrı hesapladığı için bu alanda sıklıkla kullanılmaktadır (Sachdev ve Gupta, 2019; Shi vd., 2019).

### 3.16. Performans Değerlendirme Ölçütleri

Belirli bir sorun ortaya çıktıktan sonra çözümlenme sürecinde modellemenin yanında en iyi modelin tespit edilebilirliği de modelleme kadar önemlidir. Bu sebeple modelin değerlendirilmesine ihtiyaç bulunmaktadır. Doğru bir yorumlanmanın yapılabilmesi adına performans metriklerinin hesaplanması gerekmektedir. Çeşitli tekniklerin değerlendirilmesi için en temel ölçütler aşağıda belirtilmiştir.

Doğruluk (Accuracy), doğru sınıflandırılmış örnekler ile tüm örneklerin arasındaki oranı(3.7a) temsil eder ve değerin 1'e olan yakınlığı kabul edilir.

$$Doğruluk = \frac{\text{Doğru Pozitif(TP)}+\text{Doğru Negatif(TN)}}{\text{Doğru Pozitif(TP)}+\text{Doğru Negatif(TN)}+\text{Yanlış Pozitif(FP)}+\text{Yanlış Negatif(FN)}} \quad (3.7a)$$

Kesinlik(Precision), doğru pozitif etkileşimlerin oranını(3.7b) gösteririr.

$$Kesinlik = \frac{\text{Doğru Pozitif(TP)}}{\text{Doğru Pozitif(TP)}+\text{Yanlış Pozitif(FP)}} \quad (3.7b)$$

Duyarlılık- Hassasiyet (Recall/Sensitivity), doğru şekilde tanınan pozitif etkileşimlerin(3.7c) oranını belirtir.

$$\text{Duyarlılık} = \frac{\text{Doğru Pozitif(TP)}}{\text{Doğru Pozitif(TP)}+\text{Yanlış Negatif(FN)}} \quad (3.7c)$$

Doğruluk ve geri çağırmanın harmonik ortalaması olan F1-skoru, hem hatalı pozitif hem de hatalı negatif tahminleri hesaba katarak bir sonuç ortaya çıkarır. F1 skoru İHE tahmini gibi dengesiz veri setlerine sahip modellerin değerlendirilmesinde doğruluk ve MCC gibi oldukça kullanılan metriklerden biridir (Sachdev ve Gupta, 2019).

$$\text{F1 Skoru} = 2 * \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \tag{3.7d}$$

İkili sınıflandırma problemleri için en yaygın kullanılan ölçütler arasında karmaşıklık matrisleri üzerinden hesaplanan doğruluk(accuracy) ve F1 puanı yer almaktadır. Ancak bu performans metrikleri, özellikle İHE tahmini gibi dengesiz veri kümelerinde tehlikeli bir şekilde aşırı iyimser abartılı sonuçlar ortaya çıkarabilir. Bu sebeple daha güvenilir bir performans metriği oran olan Matthews korelasyon katsayısı(MCC) da hesaplanmış ve değerlendirilmeye alınmıştır. MCC yalnızca tahmin, karmaşıklık matrisinin dört kategorisinin her birinde (doğru pozitifler(TP), yanlış negatifler(FP), doğru negatifler(TN) ve yanlış pozitifler(FP)), veri kümesinin pozitif ve negatif öğelerinin boyutuyla orantılı olarak iyi performans gösterdiğinde yüksek bir puan verir. MCC değerinin -1 ile +1 arasında ki değeri kabul edilmektedir. Değerin +1 e yakınlığı modelin güvenilirliğini arttırırken -1 e yakınlığı aralarında tamamen ters bir sınıflandırmanın olduğunu gösterir. MCC değerinin 0 olması ise rastgele bir sınıflandırma modelini belirtmektedir (Atas ve Doğan, 2022; Chicco ve Jurman, 2020).

$$MCC = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP}) * (\mathrm{TP}+\mathrm{FN}) * (\mathrm{TN}+\mathrm{FP}) * (\mathrm{TN}+\mathrm{FN})}} \tag{3.7e}$$

Bir diğer değerlendirme metriği de, ilgili tahmin edicinin farklı eşik değerlerindeki performansının gösterildiği ROC(Receiver Operating Characteristic Curve) eğrisidir. ROC eğrisi, gerçek pozitif oran ile yanlış pozitif oran sayılarının grafiğe geçirilmesiyle oluşturulur ve karşılaştırmak için de ROC Eğrisi Altındaki Alan (AUC- Area Under Curve) belirlenir. AUC ise eğri üzerindeki çeşitli noktalardaki değerlerin toplamını verir. Sonuç 0 ile 1 arasında değişmekte olup ideali 1'e yakın olanıdır (Chicco ve Jurman, 2020).

Precision- recall eğrisi ise ROC eğrisine benzer şekilde, ikili bir sınıflandırma ile elde edilen tüm olası pozitif tahmin değerlerini ve duyarlılıkları incelemek için kullanılmaktadır. Performans değerlendirilmesi olarak eğrinin altında kalan alan yine AUC'ye benzer şekilde oluşturulan eğrinin altında kalan alan (AUPR) performans değerlendirilmesi için kullanılır (Sachdev ve Gupta, 2019).

# 4. ARAŞTIRMA BULGULARI VE TARTIŞMA

Tez çalışmasında, Alzheimer hastalığına ait protein verilerinin ve ilaçların temsillerinin kullanılarak İHE tahminin yapıldığı bir keminformatik metot gerçekleştirilmiştir. Bu doğrultuda, çalışmamızda protein temsillerinin oluşturulması için Terapötik hedef veri tabanından (Y. Zhou vd, 2022) elde ettiğimiz Alzheimer ilişkili 135 proteinin bir boyutlu protein sekans verileri Uniprot (Consortium, 2019) veri tabanından indirilmiştir. Daha sonra her bir protein verisinin temsili için sekans dizileri, Word2Vec algoritması tabanlı bir teknik olan protein vektörlerine (ProtVec) dönüştürülmüştür (Asgari ve Mofrad, 2015). ProtVec, biyolojik verilerde makine öğrenmesi uygulamalarında büyük bir kolaylık sağlamaktadır. Şekil 4.1. 'de Python programı kullanılarak protein sekans dizilerinin kelimelere bir diğer adıyla 3-mer'li alt dizilere yani biyolojik adı olan kodonlara bölünmüş şekli gösterilmektedir.

Protein vektörlerinin oluşturulması yaklaşımında, kelime temsillerini bulmak amacıyla biyolojik veriler üzerinde de sıklıkla gerçekleştirilen doğal dil işleme (NLP) çalışmalarının uygulandığı Word2Vec algoritması kullanılmıştır (Mikolov vd., 2013). Bu algoritmanın sahip olduğu 2 mimariden biri olan Skip-gram mimarisi kullanılarak kelimelerin temsilleri bulunmuştur ve her bir proteinin 3-gramlı alt dizileri yani kodonları 1*100 boyutlu vektörlere dönüştürülmüştür. Bu sayede protein vektörleri elde edilmiştir. Bu temsilin avantajı, öğrenilmiş gömmelerden (embedding) oluşmasıdır. Protein vektörlerini ilk oluşturan araştırmacılar sonraki çalışmalarda daha kolay bir kullanım amacıyla Swiss-Prot veri tabanında ki 546,790 sekansı bu şekilde elde etmişlerdir. Bu çalışmada, Python aracılığıyla Alzheimer hastalığı ilişkili protein vektörleri elde edilmiş ve kayıt edilmiştir. Şekil 4.2.' de çalışılan proteinlere ait vektör gösterimleri belirtilmiştir.

```
Sample sequence
MSIIGATRLQNDKSDTYSAGPCYAGGCSAFTPRGTCGKDWDLGEQTCASGFCTSQPLCARIKKTQVCGLRYSSKGKDPLVSAEWDSRGAPYVRCTYDADLIDTQAQVDQFVSMF
GESPSLAERYCMRGVKNTAGELVSRVSSDADPAGGWCRKWYSAHRGPDQDAALGSFCIKNPGAADCKCINRASDPVYQKVKTLHAYPDQCWYVPCAADVGELKMGTQRDTPTNC
PTQVCQIVFNMLDDGSVTMDDVKNTINCDFSKYVPPPPPPKPTPPTPPTPPTPPTPPTPPTPPTPRPVHNRKVMFFVAGAVLVAILISTVRW

Convert sequence to list of kmers
[['MSI', 'IGA', 'TRL', 'QND', 'KSD', 'TYS', 'AGP', 'CYA', 'GGC', 'SAF', 'TPR', 'GTC', 'GKD', 'WDL', 'GEQ', 'TCA',
'SGF', 'CTS', 'QPL', 'CAR', 'IKK', 'TQV', 'CGL', 'RYS', 'SKG', 'KDP', 'LVS', 'AEW', 'DSR', 'GAP', 'YVR', 'CTY', 'D
AD', 'LID', 'TQA', 'QVD', 'QFV', 'SMF', 'GES', 'PSL', 'AER', 'YCM', 'RGV', 'KNT', 'AGE', 'LVS', 'RVS', 'SDA', 'DP
A', 'GGW', 'CRK', 'WYS', 'AHR', 'GPD', 'QDA', 'ALG', 'SFC', 'IKN', 'PGA', 'ADC', 'KCI', 'NRA', 'SDP', 'VYQ', 'KV
K', 'TLH', 'AYP', 'DQC', 'WYV', 'PCA', 'ADV', 'GEL', 'KMG', 'TQR', 'DTP', 'TNC', 'PTQ', 'VCQ', 'IVF', 'NML', 'DD
G', 'SVT', 'MDD', 'VKN', 'TIN', 'CDF', 'SKY', 'VPP', 'PPP', 'PKP', 'TPP', 'TPP', 'TPP', 'TPP', 'TPP', 'TPP', 'TP
P', 'TPR', 'PVH', 'NRK', 'VMF', 'FVA', 'GAV', 'LVA', 'ILI', 'STV'], ['SII', 'GAT', 'RLQ', 'NDK', 'SDT', 'YSA', 'GP
C', 'YAG', 'GCS', 'AFT', 'PRG', 'TCG', 'KDW', 'DLG', 'EQT', 'CAS', 'GFC', 'TSQ', 'PLC', 'ARI', 'KKT', 'QVC', 'GL
R', 'YSS', 'KGK', 'DPL', 'VSA', 'EWD', 'SRG', 'APY', 'VRC', 'TYD', 'ADL', 'IDT', 'QAQ', 'VDQ', 'FVS', 'MFG', 'ES
P', 'SLA', 'ERY', 'CMR', 'GVK', 'NTA', 'GEL', 'VSR', 'VSS', 'DAD', 'PAG', 'GWC', 'RKW', 'YSA', 'HRG', 'PDQ', 'DA
A', 'LGS', 'FCI', 'KNP', 'GAA', 'DCK', 'CIN', 'RAS', 'DPV', 'YQK', 'VKT', 'LHA', 'YPD', 'QCW', 'YVP', 'CAA', 'DV
G', 'ELK', 'MGT', 'QRD', 'TPT', 'NCP', 'TQV', 'CQI', 'VFN', 'MLD', 'DGS', 'VTM', 'DDV', 'KNT', 'INC', 'DFS', 'KY
V', 'PPP', 'PPP', 'KPT', 'PPT', 'PPT', 'PPT', 'PPT', 'PPT', 'PPT', 'PPT', 'PRP', 'VHN', 'RKV', 'MFF', 'VAG', 'AV
L', 'VAI', 'LIS', 'TVR'], ['IIG', 'ATR', 'LQN', 'DKS', 'DTY', 'SAG', 'PCY', 'AGG', 'CSA', 'FTP', 'RGT', 'CGK', 'DW
D', 'LGE', 'QTC', 'ASG', 'FCT', 'SQP', 'LCA', 'RIK', 'KTQ', 'VCG', 'LRY', 'SSK', 'GKD', 'PLV', 'SAE', 'WDS', 'RG
A', 'PYV', 'RCT', 'YDA', 'DLI', 'DTQ', 'AQV', 'DQF', 'VSM', 'FGE', 'SPS', 'LAE', 'RYC', 'MRG', 'VKN', 'TAG', 'EL
V', 'SRV', 'SSD', 'ADP', 'AGG', 'WCR', 'KWY', 'SAH', 'RGP', 'DQD', 'AAL', 'GSF', 'CIK', 'NPG', 'AAD', 'CKC', 'IN
R', 'ASD', 'PVY', 'QKV', 'KTL', 'HAY', 'PDQ', 'CWY', 'VPC', 'AAD', 'VGE', 'LKM', 'GTQ', 'RDT', 'PTN', 'CPT', 'QV
C', 'QIV', 'FNM', 'LDD', 'GSV', 'TMD', 'DVK', 'NTI', 'NCD', 'FSK', 'YVP', 'PPP', 'PPK', 'PTP', 'PTP', 'PTP', 'PT
P', 'PTP', 'PTP', 'PTP', 'PTP', 'RPV', 'HNR', 'KVM', 'FFV', 'AGA', 'VLV', 'AIL', 'IST', 'VRW']]
```

```
Iteration 0, loss=0.7316372990608215
Iteration 0, loss=0.6979584693908691
Iteration 0, loss=0.6028254628181458
Iteration 0, loss=0.7622859477996826
Iteration 0, loss=0.7141176462173462
Nearest to LEA: AEL, LEQ, LQA, ARA, EME, LEE, LAE, ARL,
Nearest to SAA: AAT, ASA, AAS, AAA, AAG, GAA, PAA, AAV,
Nearest to GAL: GLA, LAG, LAA, GVA, AVL, LAV, AAV, AAL,
Nearest to LLS: AIP, FWG, ITN, QMS, ATV, VQW, VEW, YIG,
Nearest to GAA: GAI, AAG, SAA, LVV, QMD, MIT, GAV, AAA,
Nearest to LKE: LEK, LKK, KKL, KEL, EKL, LQK, EEI, KLK,
Nearest to LLL: MLL, LLV, FLL, VLL, CLL, LVL, LLI, CLW,
Nearest to ALK: EAL, KAE, KEA, AKV, AEI, QAK, AIK, VLK,
Nearest to LVA: LIA, FWL, WIG, VEH, LAI, LLA, FWI, QCL,
Nearest to LEL: ELL, EEL, LEK, LKL, LEQ, EIR, LEE, VEL,
Nearest to VVA: IVA, VIA, AAI, AAD, AVT, AVV, LAA, VAA,
Nearest to LLD: RML, DVL, GHV, LLE, DFA, ILD, TLE, ELT,
Nearest to ASL: ASA, ALS, AVS, ALC, SPL, AAS, QAS, ASP,
Nearest to SSL: LSS, SSF, VSS, SFS, SSI, SQS, ASS, SNL,
Nearest to ALG: LAG, LAV, LGL, IAA, FFG, GAL, MLA, FLG,
```

Şekil 4.1. Protein sekans verilerinin Python'da 3-mer ile ayrılması ve sonrasında en yakın aminoasitlerin bulunup eğitilmesi.

```
protvec.head()
```

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ... | 90 | 91 | 92 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | -18.112476 | -23.701257 | -50.201439 | -23.098780 | -14.654442 | -36.657947 | -26.678769 | -13.642391 | -37.342159 | -33.932835 | ... | 0.471724 | -2.644732 | -1.442510 |
| **1** | -3.175629 | -7.894448 | -20.594988 | -1.664396 | -3.009061 | -0.488606 | -1.037302 | -8.812832 | -2.697608 | -5.567350 | ... | 2.430716 | -13.403044 | -2.575371 |
| **2** | -0.512001 | -2.032365 | -2.272504 | 1.048114 | -3.735968 | 0.462023 | -37.479516 | -17.867963 | -4.461757 | -0.281552 | ... | 8.530001 | 4.879894 | -1.765796 |
| **3** | -31.714723 | -23.127643 | -8.043224 | -10.526655 | -30.617627 | -29.934373 | -31.818032 | -48.695629 | -32.072113 | -29.641418 | ... | -25.539804 | -40.216376 | -37.716105 |
| **4** | -10.520748 | -23.088868 | -19.598752 | -28.398840 | -17.427234 | -6.932209 | -41.655564 | -35.856311 | -39.969317 | -30.203606 | ... | 8.105451 | -7.437048 | 1.842524 |

5 rows × 100 columns

Şekil 4.2. Protein temsili gösterimi

Çalışmada ilaç temsili için ise öncelikle farmakolojik olarak pozitif İHE verisine sahip FDA onaylı ilaçların SMILES dizileri DrugBank veri tabanından indirilmiştir. Şekil 4.3'de ilaçların SMILES dizilerine ait gösterimi belirtilmiştir.

| | DrugID | InChI | SMILES |
|---|---|---|---|
| **0** | DB00006 | InChI=1S/C98H138N24O33/c1-5-52(4)82(96(153)122... | CC[C@H](C)[C@H](NC(=O)[C@H](CCC(O)=O)NC(=O)[C@... |
| **1** | DB00007 | InChI=1S/C59H84N16O12/c1-6-63-57(86)48-14-10-2... | CCNC(=O)[C@@H]1CCCN1C(=O)[C@H](CCCNC(N)=N)NC(=... |
| **2** | DB00014 | InChI=1S/C59H84N18O14/c1-31(2)22-40(49(82)68-3... | CC(C)C[C@H](NC(=O)[C@@H](COC(C)(C)C)NC(=O)[C@H... |
| **3** | DB00027 | InChI=1S/C96H135N19O16/c1-50(2)36-71(105-79(11... | CC(C)C[C@@H](NC(=O)CNC(=O)[C@@H](NC=O)C(C)C)C(... |
| **4** | DB00035 | InChI=1S/C46H64N14O12S2/c47-35(62)15-14-29-40(... | NC(=O)CC[C@@H]1NC(=O)[C@H](CC2=CC=CC=C2)NC(=O)... |

Şekil 4.3. İlaçların SMILES dizilerinin gösterimi

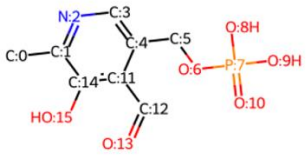
Açık kaynaklı bir Python kütüphanesi olan RDKit kütüphanesi kullanılmıştır. Bu süreçte ilaçların SMILES dizileri, Google Colab bulut sunucusu üzerinden iki boyutlu kimyasal moleküler yapılarına dönüştürülmüştür. RDKit kütüphanesi kullanılarak yazılan kodlar Şekil 4.4.'de belirtilmiştir. Oluşturulan iki boyutlu kimyasal moleküler formları, png görüntü formatında Google drive'a kayıt edilmiştir.

```
! pip install rdkit-pypi

Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheel
Collecting rdkit-pypi
  Downloading rdkit_pypi-2022.9.3-cp38-cp38-manylinux_2_17_x86_64.manylinux2014_x8
     ---------------------------------------- 29.3/29.3 MB 51.2 MB/s eta 0:00:00
Requirement already satisfied: Pillow in /usr/local/lib/python3.8/dist-packages (f
Requirement already satisfied: numpy in /usr/local/lib/python3.8/dist-packages (fr
Installing collected packages: rdkit-pypi
Successfully installed rdkit-pypi-2022.9.3
```

```
[ ] from google.colab import drive
    drive.mount('/content/drive')

    Drive already mounted at /content/drive; to attempt to forcibly remount, call driv
```

```
[ ] # import RDKit ------------------------------------------------------------
    from rdkit import Chem
    from rdkit.Chem import Draw

    # define the smiles string and covert it into a molecule sturcture ------------
    caffeine_smiles = 'CN1C=NC2=C1C(=O)N(C(=O)N2C)C'

    mol = Chem.MolFromSmiles(caffeine_smiles)

    # draw the modecule ------------------------------------------------------------
    Draw.MolToFile(mol, 'caffeine.png')

    # draw the molecule with property ----------------------------------------------
    for i, atom in enumerate(mol.GetAtoms()):
        atom.SetProp("molAtomMapNumber", str(atom.GetIdx()))
        Draw.MolToFile(mol, 'drive/MyDrive/drugimage/ caffeine_with_prop.png')
```

Şekil 4.4. Örnek bir kimyasal formun SMILES dizi temsilinin iki boyutlu kimyasal moleküler yapısına dönüştürülmesini sağlayan RDKit kütüphanesi kodları




CC(C)C[C@H](NC(=O)[C@@H](COC(C)(C)C)NC(=O)[C@H](CC1=CC=C(O)C=C1)NC(=O)[C@H](CO)NC(=O)[C@H](CC1=CNC2=CC=CC=C12)NC(=O)[C@H](CC1=CN=CN1)NC(=O)[C@@H]1CCC(=O)N1)C(=O)N[C@@H](CCCN=C(N)N)C(=O)N1CCC[C@H]1C(=O)NNC(N)=O

Şekil 4.5. Örnek olarak bir ilacın RDKit kütüphanesi kodları ile oluşturulan iki boyutlu kimyasal yapı görüntüsü ve SMILES dizisi

İlaçların kimyasal formlarına ait görüntülerden özniteliklerin çıkarılmasını sağlayan evrişimli sinir ağı mimarisinde; evrişimli katman, tam bağlı katman, havuzlama katmanı kullanılmış olup aktivasyon fonksiyonu olarak da ReLU kullanılmıştır. Böylece kayıt edilen her bir ilaç için 300 öznitelik çıkartılmıştır. Şekil 4.6.'da öznitelikleri oluşturulmuş ilaç temsilleri gösterilmektedir.

```
drug.head(10)
```

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ... | 291 | 292 | 293 | 294 | 295 | 296 | 297 | 298 | 299 | Drug |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.000000 | 0.0 | 0.0 | 17.560219 | 6.029813 | 0.0 | 0.000000 | 0.0 | 0.0 | ... | 0.0 | 30.310276 | 58.920868 | 0.000000 | 77.910290 | 0.0 | 0.0 | 0.0 | 0.0 | DB11093 |
| 1 | 0.0 | 0.000000 | 0.0 | 0.0 | 43.446995 | 44.628284 | 0.0 | 0.000000 | 0.0 | 0.0 | ... | 0.0 | 16.724203 | 88.660600 | 16.809017 | 46.856174 | 0.0 | 0.0 | 0.0 | 0.0 | DB01253 |
| 2 | 0.0 | 8.267031 | 0.0 | 0.0 | 11.746761 | 27.913773 | 0.0 | 0.000000 | 0.0 | 0.0 | ... | 0.0 | 1.707439 | 66.521590 | 0.000000 | 57.268270 | 0.0 | 0.0 | 0.0 | 0.0 | DB00862 |
| 3 | 0.0 | 0.000000 | 0.0 | 0.0 | 34.723362 | 1.697510 | 0.0 | 0.000000 | 0.0 | 0.0 | ... | 0.0 | 31.370392 | 87.435394 | 0.000000 | 60.901650 | 0.0 | 0.0 | 0.0 | 0.0 | DB00887 |
| 4 | 0.0 | 12.395467 | 0.0 | 0.0 | 12.449865 | 51.583250 | 0.0 | 0.000000 | 0.0 | 0.0 | ... | 0.0 | 6.196685 | 68.661490 | 0.000000 | 66.700035 | 0.0 | 0.0 | 0.0 | 0.0 | DB09027 |
| 5 | 0.0 | 7.951857 | 0.0 | 0.0 | 47.738850 | 58.519573 | 0.0 | 2.842525 | 0.0 | 0.0 | ... | 0.0 | 14.035389 | 41.613830 | 1.826044 | 72.386330 | 0.0 | 0.0 | 0.0 | 0.0 | DB12407 |
| 6 | 0.0 | 7.512576 | 0.0 | 0.0 | 36.740005 | 28.680279 | 0.0 | 0.000000 | 0.0 | 0.0 | ... | 0.0 | 22.515741 | 83.812560 | 0.000000 | 48.407246 | 0.0 | 0.0 | 0.0 | 0.0 | DB00357 |
| 7 | 0.0 | 0.000000 | 0.0 | 0.0 | 33.749924 | 44.191190 | 0.0 | 5.723521 | 0.0 | 0.0 | ... | 0.0 | 22.981112 | 51.613426 | 0.000000 | 53.404340 | 0.0 | 0.0 | 0.0 | 0.0 | DB00562 |
| 8 | 0.0 | 0.000000 | 0.0 | 0.0 | 18.808336 | 20.519950 | 0.0 | 0.000000 | 0.0 | 0.0 | ... | 0.0 | 29.096786 | 67.745360 | 0.000000 | 57.741547 | 0.0 | 0.0 | 0.0 | 0.0 | DB13211 |
| 9 | 0.0 | 3.067825 | 0.0 | 0.0 | 0.000000 | 45.588985 | 0.0 | 0.000000 | 0.0 | 0.0 | ... | 0.0 | 10.432179 | 87.583770 | 0.000000 | 61.558933 | 0.0 | 0.0 | 0.0 | 0.0 | DB00929 |

10 rows × 301 columns

Şekil 4.6. Özellik çıkarımı oluşturulmuş ilaçlara bir örnek

Biyolojik ve kimyasal veriler üzerinde çalışıldığında, öznitelik çıkarımı eğitim verilerinin kalitesi üzerinde bir etkiye sahiptir. Araştırmalarda, öznitelik çıkarımı ile ayrıntılı ilaç temsilleri oluşturulmaktadır. Öznitelik çıkarımı ile Transformers ağlarının kullanımı İHE araştırma ve geliştirmesini ilerletmek adına, ilaç molekülü yapısındaki özellik bilgilerini sağladığı için avantajlıdır. İlaçların iki boyutlu yapılarından elde edilen görüntüleri Transformers ağı ile öznitelik çıkarımı için Google Colab ortamında çalışılmıştır. Öncelikle görüntüler png formatından RCB formatına dönüştürülmüş, sonrasında Python yazılım programı üzerinde Transformers ağı aracılığıyla görüntü öznitelikleri çıkarılmıştır. Şekil 4.7. 'de Transformers ağı kullanılarak çıkarılan öznitelikler gösterilmektedir.

```
drug.head(10)
```

| | Drug | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | DB06704 | -0.417919 | 0.074360 | 0.001515 | 1.536581 | -0.665466 | 1.096165 | 0.420074 | -0.973725 | 0 |
| 1 | DB06705 | 0.795331 | 0.607815 | 0.315204 | 0.402940 | 0.257041 | -0.296228 | 0.095356 | -0.987730 | -0 |
| 2 | DB06706 | -0.131663 | -0.483524 | 0.133624 | 1.129180 | 0.353919 | -0.183939 | 0.494740 | -0.257505 | 0 |
| 3 | DB06707 | 0.228828 | -0.668520 | -0.125668 | 1.525549 | 0.125082 | 0.554882 | 0.514334 | -0.626240 | 0 |
| 4 | DB06711 | -0.216400 | -1.074676 | -0.237418 | 0.865100 | 0.468219 | 0.317024 | 0.300193 | -0.380998 | 0 |
| 5 | DB06709 | 0.109067 | -0.348653 | 0.428121 | 1.279374 | -0.172823 | 0.540607 | 1.320521 | -0.977919 | 0 |
| 6 | DB06708 | -0.688770 | -0.059466 | 0.035469 | 0.154409 | 1.002248 | 1.196975 | 0.188385 | -0.216616 | 0 |
| 7 | DB06710 | -0.495777 | 0.411862 | -0.312100 | 1.024052 | -0.149665 | 0.968492 | -0.258474 | -0.379308 | 0 |
| 8 | DB06713 | 0.898513 | 0.517110 | 0.441484 | 0.734006 | 0.392594 | -0.518450 | 0.459279 | -0.511764 | 0 |
| 9 | DB06712 | 0.270630 | -0.961489 | -0.370217 | 1.245387 | 1.259165 | 0.661886 | 0.026870 | -0.844592 | -0 |

10 rows × 301 columns

Şekil 4.7. Transformers ağı kullanılan ilaçların öznitelikleri

İHE çiftlerinde negatif etiketli veriler uzaklık yöntemleri kullanılarak oluşturulmuştur. Öncelikle protein vektörleri ile elde edilen hedef verisi üzerinde, her hedef için en uzak protein; Öklid, Manhattan ve Minkowski uzaklık yöntemleri aracılığı ile belirlenmiştir. Bunun sonucunda her proteinin en uzak olduğu hedefin ilaçları negatif/ bilinmeyen etkileşim verisi olarak kabul edilmiştir ve 0 ile etiketlenmiştir. Rastgele oluşturulan negatif etiketli ilaçlar ile birlikte toplamda İHE negatif veri çiftleri seçimi için 4 farklı yöntem kullanılarak belirlenmiştir. İHE tahmini alanında yapılan çalışmalar da sıklıkla negatif etiketli ilaçlar rastgele seçilmektedir. Bu yöntem uzun süreçte sorun oluşturabilmektedir. Bu sebeple tez çalışmasında, ilaç hedef etkileşim verisine ait negatif etiketli veriler uzaklık yöntemleri ile belirlenmiştir.

Tez çalışmasında, ilaç temsili için CNN ve transformers ağları ile oluşturulan 300 boyutlu özniteliklerin çıkarıldığı veri setleri ayrı ayrı karşılaştırılmak üzere kullanılmıştır. Hedef temsili için, öznitelik çıkarımı yapılan ve 100 boyutlu oluşturulan protein vektörleri kullanılmıştır. Hedef ve ilaç temsilleri, oluşturulan negatif ilaç hedef etkileşimi çiftleri ile birleştirilmiştir. Transformers ve CNN ağı kullanılan ilaç temsillerine ait öznitelikleri, negatif etkileşim çiftleri Öklid, Manhattan, Minkowski ve rastgele oluşturulan negatif İHE çifti değerleri ile birleştirilmiştir. Transformers ve CNN ağı ilaç temsilleri veri setleri için ayrı ayrı 4 negatif İHE çiftleri ile birleştirilmiştir. Böylece Transformers veri seti ve CNN veri setleri oluşturulmuştur.

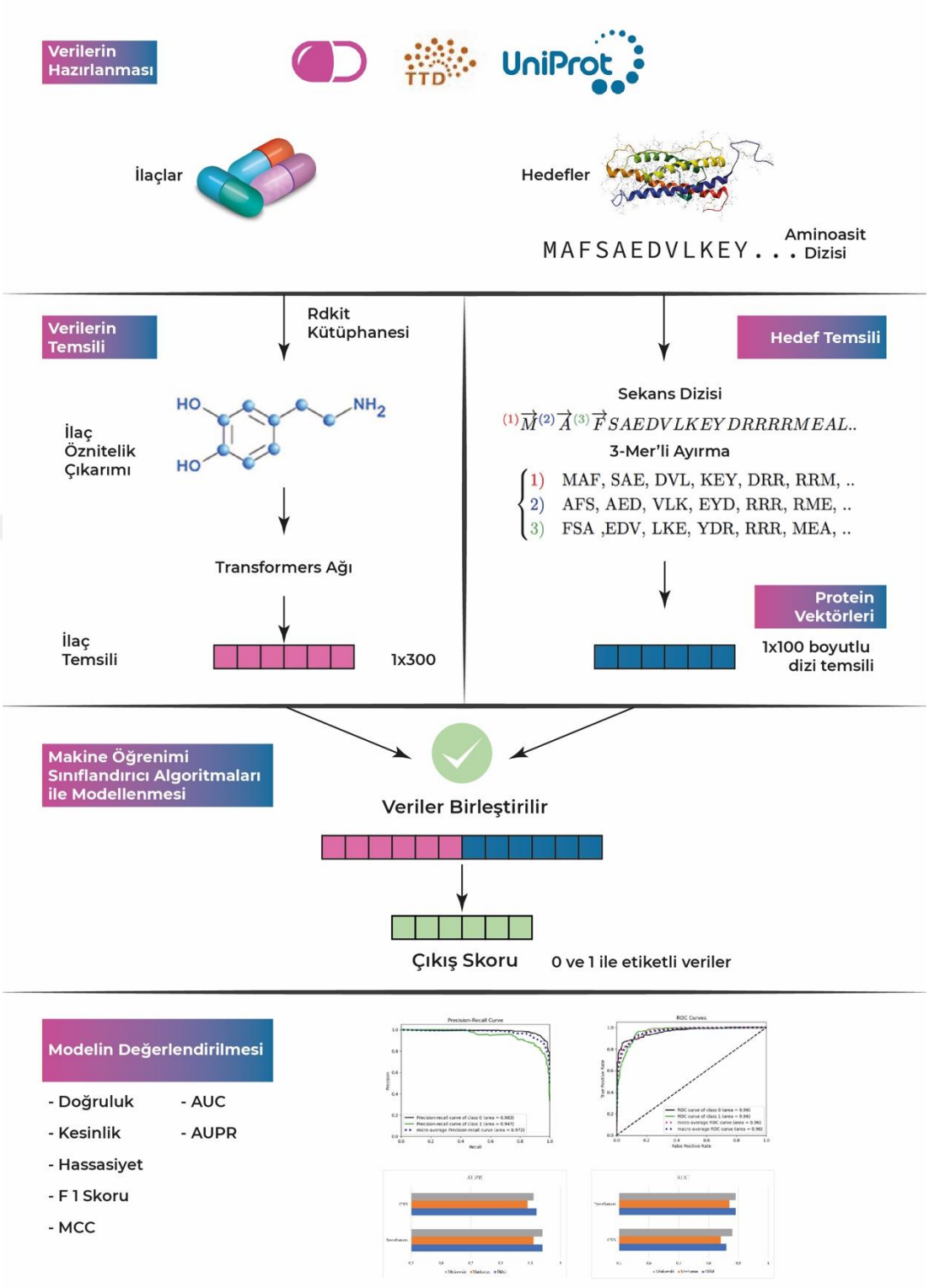
Oluşturulan veri setleri sınıflandırıcı algoritmalar olan RO, KA, LR ve DVM ile modellenmiştir. Modelleme sonucunda her veri setine ait; performans metrik değerleri, AUC ve AUPR skorları birbirleriyle karşılaştırılmıştır.

## 4.1. Oluşturulan Modellerin Performans Metrikleri

Makine öğrenmesi için hazırlanan giriş ve çıkış verileri sınıflandırıcı modellere uygulanmıştır. Bu bölümde, uygulanan sınıflandırıcı modellerin performans metrikleri yer almaktadır. İlaç temsili için uygulanan 2 farklı yöntemin sonuçları ayrı çizelgelerde belirtilmiştir. İlaçların iki boyutlu yapılarının CNN ile özelliklerinin çıkarılıp hedef ve etkileşim verisi ile birleştirildiği sınıflandırma modellerinin uygulamalarından elde edilen sonuçlar Çizelge 4.1.'de yer almaktadır.

Çizelge 4.1.'e göre İHE verisinde uygulanan Öklid uzaklık yönteminin performans metriklerinin diğer uzaklık yöntemlerine göre daha yüksek olduğu görülmektedir. Çizelge 4.1.'e göre rastgele üretilen veri setine ait değerler çizelgenin geneline oranla en düşük değerlere sahiptir. Sonuçlarımız negatif etiketli verilerin seçiminin, modellerin tahmin sonucunda ne kadar önemli olduğunu ve ne ölçüde etkilediğini açıkça ortaya koymaktadır. İHE tahmini alanı için bilinmeyen etkileşim verisinin rastgele üretilmesi yerine, çeşitli girdi verisi temsillerinin acilen modeller için işlenmesi ve uygulanmaya başlanması gerekmektedir.

Şekil 4.8'de tezin genel akış şeması bulunmaktadır. Açıklanan yöntemler ışığında, Şekil 4.8'de belirtildiği üzere tez çalışması adım adım gerçekleştirilmiştir.




Şekil 4.8. Tez çalışmasının genel akışı

Çizelge 4.1. CNN ağı veri setine ait performans metrikleri

| Negatif İHE Çifti Veri Seçim Yöntemi | Sınıflandırma Modelleri | Doğruluk % | Kesinlik % | Hassasiyet % | F1-Skor % | MCC |
|---|---|---|---|---|---|---|
| **Öklid** | **RO** | **0.8943** | **0.9394** | **0.9060** | **0.9224** | **0.7581** |
| | **LR** | 0.7940 | 0.8921 | 0.8168 | 0.8528 | 0.6581 |
| | **KA** | 0.8239 | 0.8947 | 0.8500 | 0.8717 | 0.5934 |
| | **DVM** | 0.8380 | 0.9210 | 0.8495 | 0.8838 | 0.6233 |
| **Manhattan** | **RO** | **0.8609** | **0.9184** | **0.8790** | **0.8983** | **0.6802** |
| | **LR** | 0.7852 | 0.8710 | 0.8193 | 0.8443 | 0.5012 |
| | **KA** | 0.8133 | 0.8894 | 0.8407 | 0.8644 | 0.5680 |
| | **DVM** | 0.8116 | 0.9052 | 0.8289 | 0.8654 | 0.5596 |
| **Minkowski** | **RO** | **0.8873** | **0.9342** | **0.9010** | **0.9173** | **0.7418** |
| | **LR** | 0.7922 | 0.8868 | 0.8179 | 0.8510 | 0.5143 |
| | **KA** | 0.8503 | 0.9157 | 0.8678 | 0.8911 | 0.6547 |
| | **DVM** | 0.8380 | 0.9210 | 0.8495 | 0.8838 | 0.6233 |
| **Rastgele Seçilen** | **RO** | **0.6283** | **0.9151** | 0.6596 | **0.7666** | 0.0569 |
| | **LR** | 0.6159 | 0.9045 | 0.6532 | 0.7586 | 0.0468 |
| | **KA** | 0.5274 | 0.6100 | **0.6571** | 0.6327 | 0.0273 |
| | **DVM** | 0.6212 | 0.9071 | 0.6564 | 0.7616 | **0.0790** |

İlaçların transformers ağı ile özniteliklerinin çıkarıldığı, hedef ve etkileşim verisi ile birleştirilerek makine öğrenmesi algoritmaları performans metrikleri ise Çizelge 4.2.' de yer almaktadır.

Çizelge 4.1 ve Çizelge 4.2.' de İHE verisinde, Öklid, Manhattan, Minkowski uzaklık yöntemleriyle elde edilen verilerden oluşturulan farklı veri setleri ve rastgele üretilen veri seti ayrı satırlarda belirtilmiştir. Her veri seti için RO, LR, KA ve DVM sınıflandırma modellerinin sonuçları ayrı çizelgelerde gösterilmiştir.

Çizelge 4.1.'e göre sınıflandırma modellerinin tahmin sonuçlarına bakıldığında yine çizelge genelinde RO modelinin en yüksek değerlere sahip olduğu görülmektedir. En iyi değerlendirme sonucu seçilirken sadece doğruluk(accuracy) değeri gibi tek başına bir metrik almak yerine F1 skoru ve MCC skorları ile birlikte tartışmak daha doğru sonuçlar oluşturacaktır. Bu bağlamda, en iyi tahminin RO'ya ait olduğu görülmektedir. Çizelge 4.1.'de İHE verisi rastgele oluşturulan veri setinde tüm sınıflandırma modelleri için MCC değerlerinin sıfır olduğu görülmektedir. Bu değer rastgele oluşturulan veriler için doğru olmakla birlikte diğer verilerin MCC değerinin de gerçekliğini arttırmaktadır.

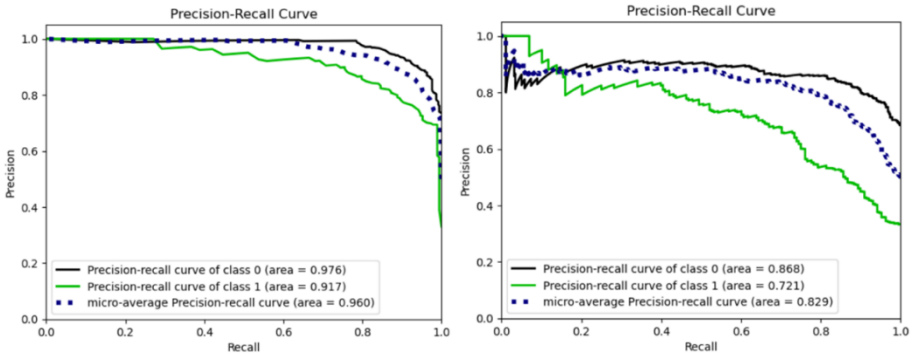
Çizelge 4.2.'de ise İHE verisi için kullanılan yöntemlerden biri olan Öklid uzaklık veri setinin diğer yöntemlerden daha iyi performans gösterdiği görülmektedir. Genel olarak çizelgede yine RO modelinin en iyi değerlere sahip olduğu gözlemlenmektedir.

Çizelge 4.1. ve Çizelge 4.2. arasında genel bir karşılaştırma yapıldığında, ilaçların temsilleri için kullanılan Transformers ağının diğer yöntemden daha iyi sonuç gösterdiği açıkça ortaya konulmuştur. Bunun sebebi ise transformers ağı ile oluşturulan temsillerin, ilaçların doğada ki formlarına benzerliğinden gelmektedir. İki çizelgenin geneline bakıldığında en yüksek değerlerin RO sınıflandırma modeline ait olması beklenilen bir durumdur. Çünkü, İHE tahmininde RO sıkça kullanılan ve iyi sonuç gösteren bir modeldir.

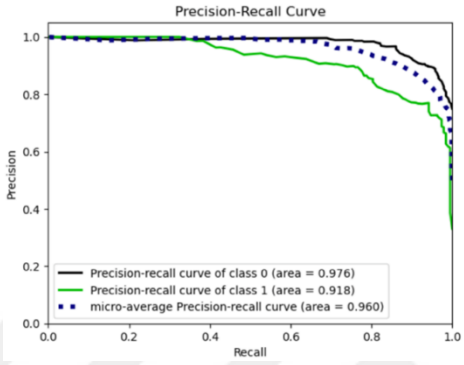
Çizelge 4.2. Transformers ağı veri setine ait performans metrikleri

| Negatif İHE Çifti Veri Seçim Yöntemi | Sınıflandırma Modelleri | Doğruluk % | Kesinlik % | Hassasiyet % | F1-Skor % | MCC |
|---|---|---|---|---|---|---|
| **Öklid** | **RO** | **0.9119** | **0.9552** | **0.9166** | **0.9355** | **0.7985** |
| | **LR** | 0.7359 | 0.8842 | 0.7601 | 0.8175 | 0.6945 |
| | **KA** | 0.8732 | 0.9315 | 0.8850 | 0.9076 | 0.7082 |
| | **DVM** | 0.8485 | 0.9289 | 0.8567 | 0.8914 | 0.6485 |
| **Manhattan** | **RO** | **0.8890** | **0.9236** | **0.9116** | **0.9176** | **0.7480** |
| | **LR** | 0.7852 | 0.8710 | 0.8193 | 0.8443 | 0.6204 |
| | **KA** | 0.8591 | 0.9236 | 0.8731 | 0.897 | 0.6750 |
| | **DVM** | 0.8450 | 0.9236 | 0.8560 | 0.8886 | 0.6404 |
| **Minkowski** | **RO** | **0.9066** | **0.9578** | **0.9077** | **0.9321** | **0.7860** |
| | **LR** | 0.7359 | 0.8842 | 0.7601 | 0.8175 | 0.7860 |
| | **KA** | 0.8820 | 0.9421 | 0.8883 | 0.9144 | 0.7284 |
| | **DVM** | 0.8914 | 0.9289 | 0.8567 | 0.8914 | 0.6485 |
| **Rastgele Seçilen** | **RO** | 0.6530 | 0.9071 | **0.6799** | 0.7772 | **0.0765** |
| | **LR** | 0.6460 | 0.9655 | 0.6606 | 0.7844 | 0.0765 |
| | **KA** | 0.5805 | 0.6790 | 0.6881 | 0.6835 | 0.0616 |
| | **DVM** | **0.6584** | **0.9867** | 0.6642 | **0.7940** | 0.0667 |

İlaç temsili için uygulanan 2 farklı yöntemden elde edilen ve etkileşim verilerinde 0 ile etiketlenen verilerin yönteminde kullanılan Öklid, Minkowski, Manhattan verilerinin tahmin sonuçlarına ait ROC eğrileri ve Precision-Recall eğrileri ayrı şekillerde belirtilmiştir.

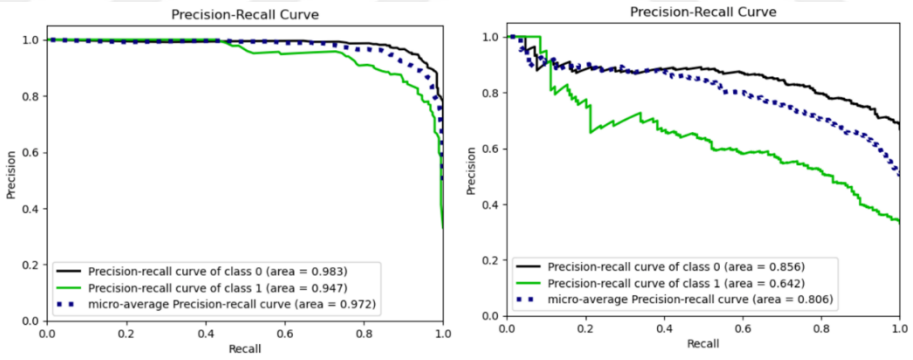



**a: Karar Ağaçları** **b: Lojistik Regresyon**

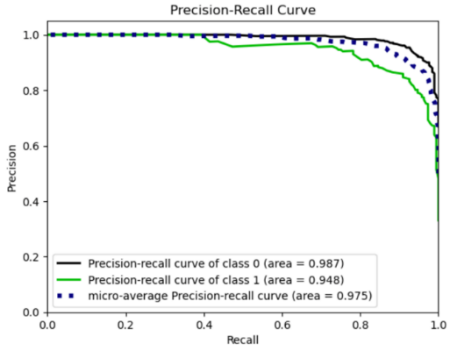



**c: Rastgele Orman**

Şekil 4.9. İHE negatif veri seçiminde Öklid yöntemi kullanılan CNN ağı veri setine ait Precision-Recall eğrileri



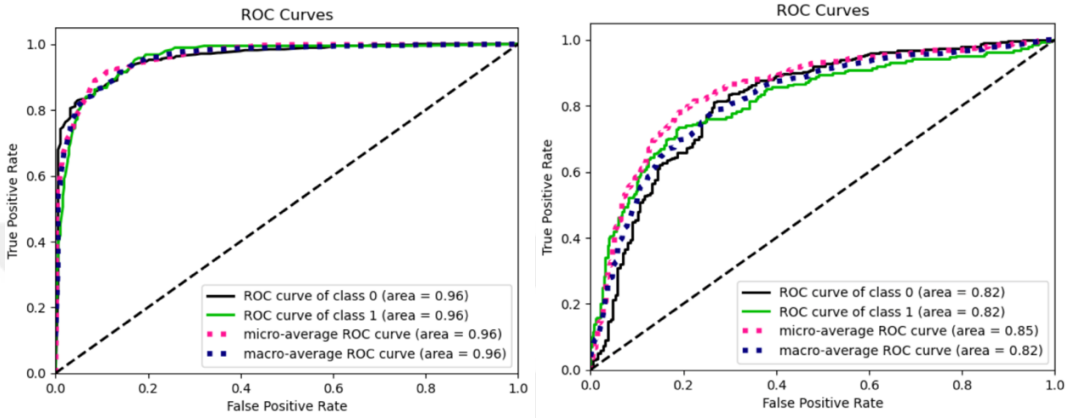

**x: Karar Ağaçları** **y: Lojistik Regresyon**




**z: Rastgele Orman**

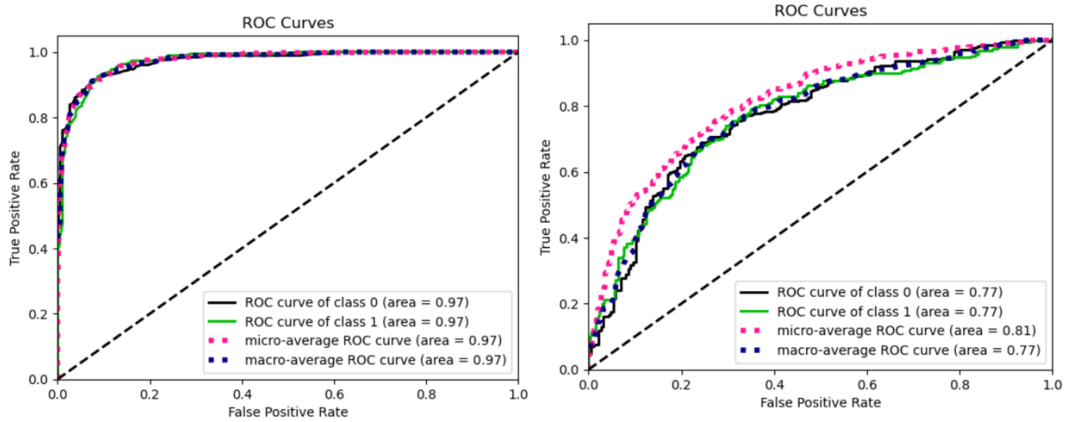
Şekil 4.10. İHE negatif veri seçiminde Öklid yöntemi kullanılan Transformers ağı veri setine ait Precision- Recall Eğrileri

Şekil 4.9. ve Şekil 4.10.'a göre en iyi sonuç veren model RO modelidir. Roc eğrisine göre Precision-Recall eğrisinin çalışmamızda kullanılmasının öncelikli sebebi ise İHE tahmini gibi veri setinin çok düzensiz olduğu çalışma alanlarında daha doğru sonuçlar elde etmesidir. Bu bağlamda bakıldığında yine transformers ağının uygulandığı veri sonuçlarına göre Öklid uzaklık yönteminde RO modeli %98.87 değeri ile en iyi değerlere sahiptir.




**a: Rastgele Orman** **b: Lojistik Regresyon**

Şekil 4.11. İHE negatif veri seçiminde Öklid yöntemi kullanılan CNN ağı veri setine ait ROC eğrileri

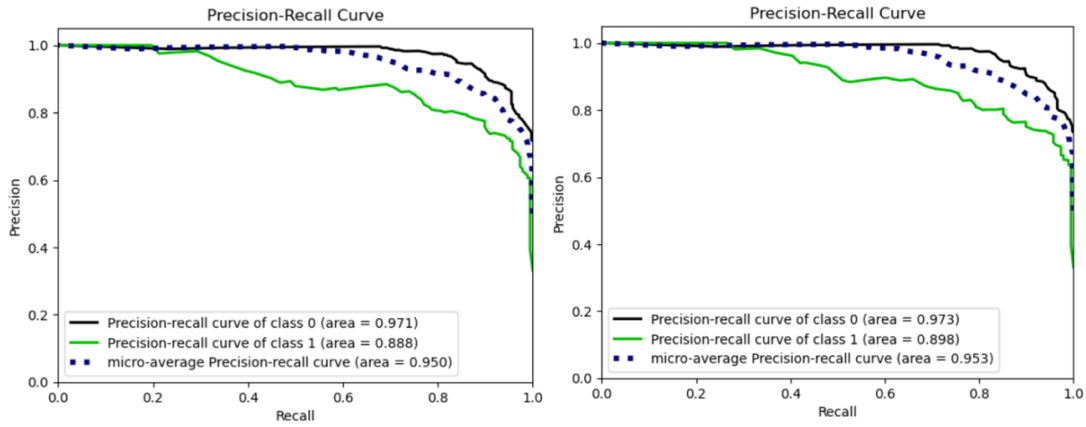



**x: Rastgele Orman** **y: Lojistik Regresyon**

Şekil 4.12. İHE negatif veri seçiminde Öklid yöntemi kullanılan Transformers ağı veri setine ait ROC eğrileri

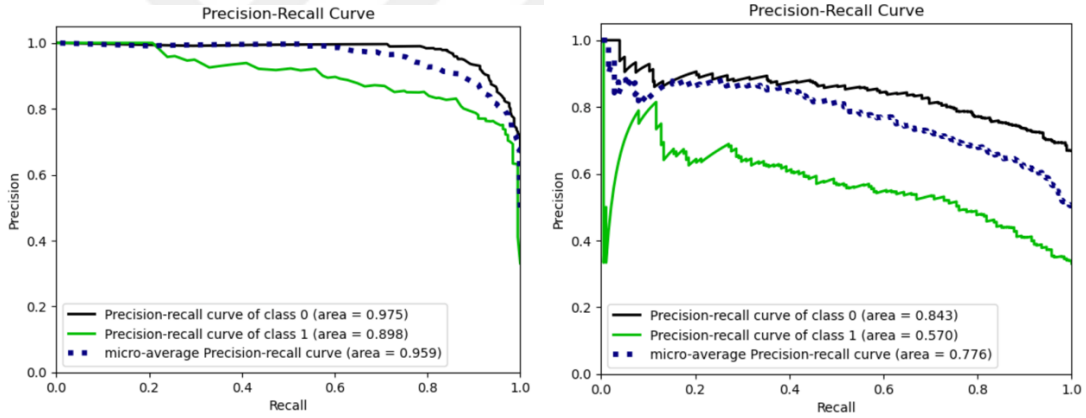
Şekil 4.11 ve Şekil 4.12 'de Öklid yöntemine göre ilaçların 2 farklı yönteminin uygulandığı modellerin sonuçlarının Roc eğrileri bulunmaktadır. Bu sonuçlar neticesinde 0.97 değeri ile transformers ağına ait RO modelinin en iyi ROC eğrisi

olduğu gösterse de LR'a ait eğri sonuçları da karşılaştırılabilir nitelikte iyi performans göstermektedir.




**a:Lojistik Regresyon** **b: Rastgele Orman**

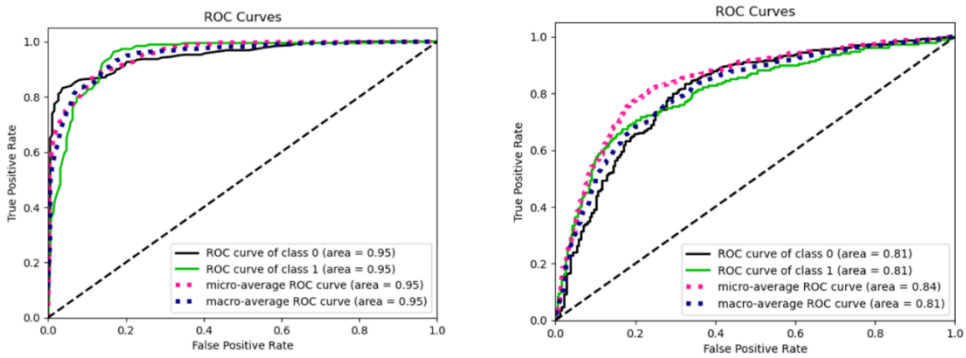
Şekil 4.13. İHE negatif veri seçiminde Manhattan yöntemine kullanılan CNN ağı veri setine ait Precision-Recall eğrileri




**x: Rastgele Orman** **y: Lojistik Regresyon**

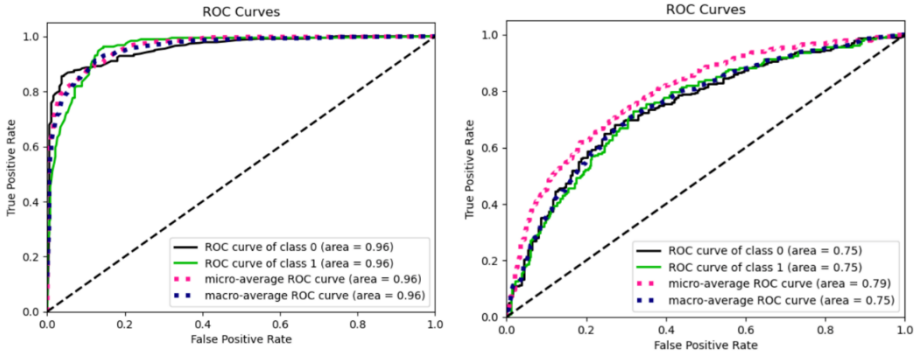
Şekil 4.14. İHE negatif veri seçiminde Manhattan yöntemi kullanılan Transformers ağı veri setine ait Precision-Recall eğrileri

Şekil 4.13. ve Şekil 4.14.'e göre en iyi sonuç veren model RO'dur. Sadece precision-recall eğrilerine bakıldığında 0.975 değeri yanıltıcı olabilmekte ve aşırı iyimser ve aşırı öğrenmeye odaklı bir sonuç olarak değerlendirilebilir. Fakat performans metriklerinin değerlendirmeye alınması ile daha sağlıklı yorumlama yapılabilmektedir. Transformers ağı için doğruluk(accuracy) değerinin %88.90, F1-Skorunun %91.76 olduğu göz önüne alındığında modelin dengeli bir sonucunun olduğu açıklanabilir.




**a: Rastgele Orman** **b: Lojistik Regresyon**

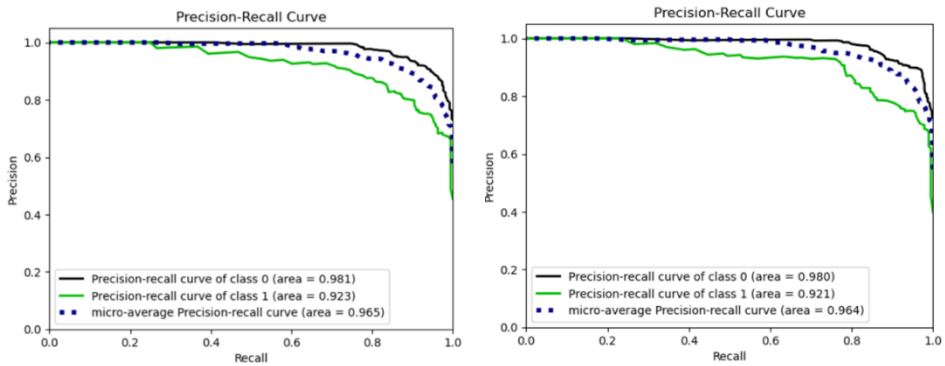
Şekil 4.15. İHE negatif veri seçiminde Manhattan yöntemi kullanılan CNN ağı veri setine ait ROC eğrileri




**x: Rastgele Orman** **y: Lojistik Regresyon**

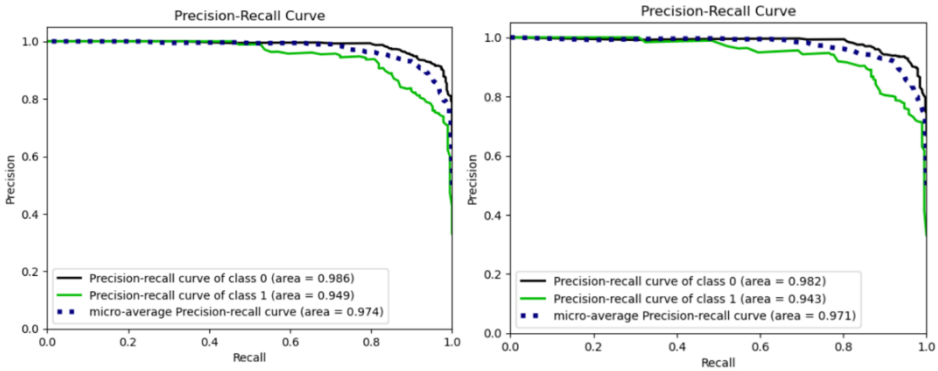
Şekil 4.16. İHE negatif veri seçiminde Manhattan yöntemi kullanılan Transformers ağı veri setine ait ROC eğrileri

Şekil 4.15. ve Şekil 4.16.'ya göre en iyi sonuç veren model RO modelidir. Bu bağlamda bakıldığında yine transformers ağının uygulandığı veri sonuçlarına göre Manhattan uzaklık yönteminde RO modeli 0.96 değeri ile en iyi değerlere sahiptir.




**a: Rastgele Orman** **b: Lojistik Regresyon**

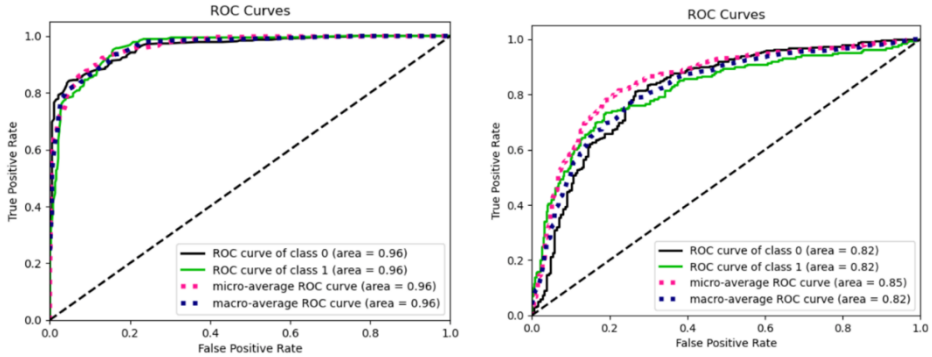
Şekil 4.17. İHE negatif veri seçiminde Minkowski yöntemi kullanılan CNN ağı veri setine ait Precision-Recall eğrileri




**x: Rastgele Orman** **y: Lojistik Regresyon**

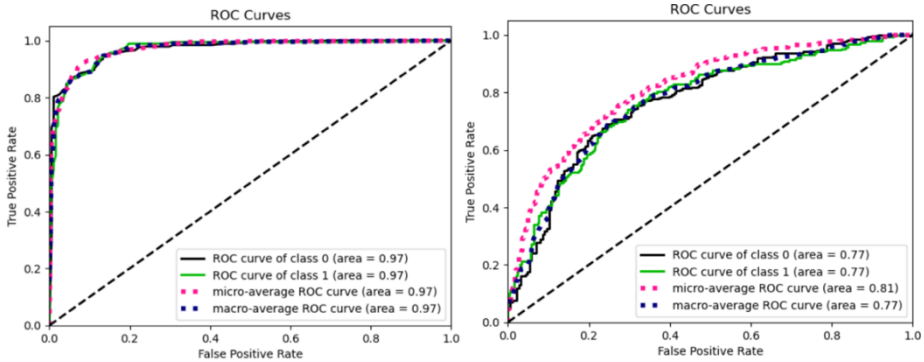
Şekil 4.18. İHE negatif veri seçiminde Minkowski yöntemi kullanılan Transformers ağı veri setine ait Precision-Recall eğrileri

Şekil 4.17. ve Şekil 4.18.'de Minkowski yöntemine ait precision- recall eğrileri bulunmaktadır. Eğrilere göre en iyi sonuç veren yöntem transformers ağına aittir ve en iyi sonuca sahip model RO modelidir.




**a: Rastgele Orman** **b: Lojistik Regresyon**

Şekil 4.19. İHE negatif veri seçiminde Minkowski yöntemi kullanılan CNN ağı veri setine ait ROC eğrileri




**x: Rastgele Orman** **y: Lojistik Regresyon**

Şekil 4.20. İHE negatif veri seçiminde Minkowski yöntemi kullanılan Transformers ağı veri setine ait ROC eğrileri

Şekil 4.19. ve Şekil 4.20.'da ilaç hedef etkileşim verisi için Minkowski yöntemine ait ROC eğrileri belirtilmiştir. Şekil 4.18. ve Şekil 4.19.'a göre en iyi sonuç veren model RO'dur. Bu bağlamda bakıldığında yine transformers ağının uygulandığı veri sonuçlarına göre Minkowski uzaklık yönteminde RO modeli 0.97 sonucu ile en iyi değere sahiptir.

Protein temsilinde, aminoasit dizilerinin sayısal vektörlere dönüştürülmesi sonucuyla oluşan özniteliklerin çıkarılması İHE tahmini için önemlidir. Çünkü, protein temsili yöntemlerinde ProtVec gibi derin öğrenme ile eğitilmiş temsiller İHE tahmininde sıkça kullanılmaktadır ve iyi performans metriklerine sahiptir. Dolayısıyla bu temsillerin makine öğrenmesi sınıflandırıcı algoritmaları ile modellenmesinin avantajları bulunmaktadır. Eğitilen verilerin tekrar bir karmaşık derin öğrenme yöntemleri ile oluşacak tahmin ve performans metrik değerleri yanlış sonuçlara neden olabilmektedir (Atas ve Doğan, 2022). Bu bağlamda, çalışmamızda verilerin temsillerinin oluşturulduktan sonra makine öğrenmesi sınıflandırma modellerine ait performans metrikleri sonuçları tartışılmıştır.

Dtigems+ araştırmasında 0.92 ile yüksek bir AUPR sonucu elde ederek yanlış pozitif oranlarını da düşürdüklerini göstermişlerdir. Fakat bu yöntemin en önemli dezavantajı, kullanımı son yıllarda sıkça tercih edilmeyen benzerlik tabanlı tekniklerin kullanılmasıdır. Çünkü ilaçların birbirine kimyasal benzerliği veya proteinlerin birbirine benzerliği aynı ilaç veya hedefe etki edeceklerinin kanıtı değil aksine hatalı sonuçlara da yol açabilmektedir. Bu sebeple, tez çalışmasında ilaç, hedef ve etkileşim veri setlerinin birlikte uygulandığı kemogenomik yöntemler ile çalışılmıştır (Thafar vd., 2020).

DeepConv-DTI araştırmasında kullanılan protein temsili Alzheimer hastalığı gibi lokal kalıntıların fazla olduğu hastalıklar için uygun olsa da bu tahmin modeli oldukça karmaşık bir modele sahiptir. Bu sonuçtan yola çıkılarak çalışmamızda ProtVec temsilinin kullanılmasının ayrıca etkisi olmuştur (Lee vd., 2019).

DeepFusion'da, kullanılan veri seti %100, %70, %50 ve %30 oranın da 4 veri setine bölünmüş ve İHE tahmini için kullanılan büyük verinin azaltıldığında da sonuçların iyi olabileceğini göstermişlerdir (Song vd., 2022). Tez kapsamında da, büyük veri

yerine hastalık bazlı İHE tahmini ile çalışılmıştır. Çalışma bağlamında İHE tahmini için büyük veri setlerinin kullanılmasının yanında hastalık bazlı uygulamalarda da iyi sonuçlar elde edilebileceğini göstermiş bulunmaktayız. Ve bunun yanı sıra yeni ilaç ve hedef temsillerine ihtiyacın gerekliliğinden dolayı, CNN ile ilaç kimyasal formlarının görüntülerinden çıkarılan öznitelikler ile transformers ağı uygulanarak elde edilen öznitelikler ile yapılan tahmin sonuçları karşılaştırılmıştır. Sonuçta, transformers ağının daha iyi sonuçlar elde ettiğini göstermiş bulunmaktayız. Bu sebeple, İHE alanı için makine öğrenmesi ve yapay zeka çağı ile bütünleşik ilerleyen çalışmalarda uygulanan klasik yöntemlerin dışına çıkılması gerekmektedir. İlaç temsili için klasikleşen moleküler parmak izi uygulamalarının yerine transformers ağı gibi yeni yöntemlerin kullanıldığı çalışmaların daha iyi sonuçların elde edilmesi İHE tahmini gibi yeni çalışma alanlarında dahi gelişime gidilmesinin en önemli örneklerinden biridir.

GraphDTI araştırmasının en önemli avantajı, heterojen verinin büyüklüğünü daha da büyütmemek adına kaçınılan yöntemlerde GraphDTI'nin verilerine eklediği gen ekspresyonu ve protein-protein etkileşim ağının birleştirilmesidir (G. Liu vd., 2021). Ve bu yöntem sonucunda da 0.996 gibi yüksek bir AUC sonucuyla da uygulanabilirliğini arttırmaktadır. İleri ki çalışmalar için hastalık bazlı tahminin yanında aynı hastalığın gen ekspresyonu ve protein protein etkileşim ağlarının da eklenmesi daha doğru sonuçların önünü açabilecektir.

Park ve Marcotte, (2012) çalışmalarında belirtildiği üzere İHE alanında eğitim ve test veri seti 4 grupta şu şekilde incelenebilir; test ve eğitim veri seti içinde ortak olarak hem ilaç hem hedef verisi ile çalışılan, sadece ilaç verisi ile çalışılan, sadece hedef verisi ile çalışılan ve ne ilaç verisinin kullanıldığı ne de hedef verisinin kullanıldığı veri setleridir. Ancak bu alan için çalışmaların performansını en yüksek seviyede etkileyen grup ve ayrıca tez çalışmasında da çalışılan grup olan, hem ilaç hem de hedef verisinin kullanıldığı veri setidir (Pahikkala vd., 2015; Park ve Marcotte, 2012).

İHE tahmini alanında makine öğrenmesi tabanlı modellerde, performans metrik değerlerini büyük ölçüde temsiller için uygulanan farklı yöntemlerin ve negatif etkileşim veri setinin önemi yadsınamazdır. Bu sebeple çalışmamızda İHE dışı verilerin haricinde kalan uzay alanında ki ilaçları farklı uzaklık yöntemleri ile seçilen

İHE çiftleri ile ilerlenmiştir. Çalışmamızda uzaklık yöntemi uygulanan veri setlerinin rastgele seçilenler ile karşılaştırıldığında daha iyi sonuçlar elde edildiği gözlemlenmiştir. Diğer bir yönden, ilaç temsili için CNN ağı ile karşılaştırıldığında, transformers ağının kullanıldığı veri setlerinde ki performans metrik değerlerinin artışı girdi verilerine ait temsillerin ve uygulanan yöntemlerin geliştirilmesinin önemini de açıkça ortaya koymaktadır.

## 5. SONUÇ VE ÖNERİLER

İlaç keşfinin temel ve ehemmiyetli adımlarından biri olan ilaç hedef etkileşimlerinin (İHE) öngörülmesi ilaç sanayisinin yanı sıra hastaların iyileşme süreçlerinin kısalması için de oldukça önemlidir (Sachdev ve Gupta, 2019). Bir hastalığın doğru İHE'lerinin bulunması o hastalığın tedavi yöntemlerinin güvenilir olmasını belirler. Bu sebeple ilaç keşfi alanında yeni İHE'nin belirlenmesini kolaylaştırmak amacıyla makine öğrenime algoritmalarının ve hesaplamalı yöntemlerin kullanımı yaygınlaşmıştır (Li vd., 2023).

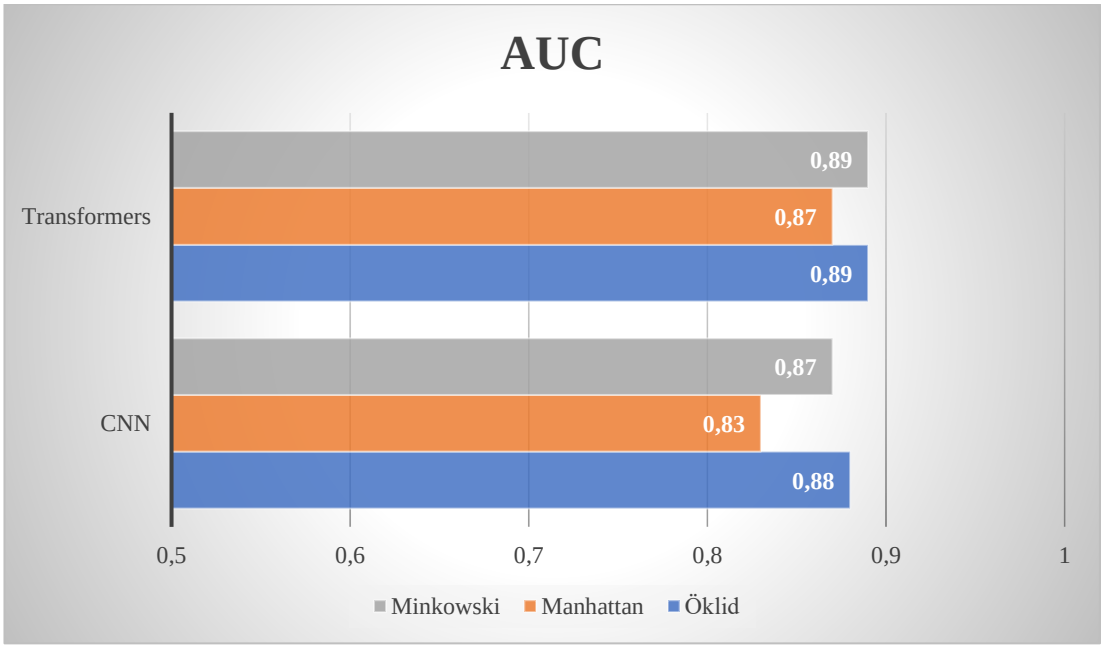
Çizelge 5.1. İlaçların CNN ve Transformers veri setlerine ait AUC - AUPR değerleri

| İHE Çifti İçin Veri Seçimi | Sınıflandırma Modelleri | CNN AUC | Transformers AUC | CNN AUPR | Transformers AUPR |
|---|---|---|---|---|---|
| **Öklid** | **RO** | 0.8726 | **0.8898** | 0.9232 | **0.9424** |
| | **LR** | 0.8180 | **0.8898** | 0.7182 | 0.6416 |
| | **KA** | 0.8117 | **0.8354** | 0.6538 | **0.9424** |
| | **DVM** | 0.7956 | **0.8075** | **0.9235** | 0.7091 |
| **Manhattan** | **RO** | 0.8381 | **0.8714** | 0.8960 | **0.9071** |
| | **LR** | **0.8061** | 0.7451 | **0.6883** | 0.5699 |
| | **KA** | 0.7745 | **0.8103** | 0.6204 | **0.6602** |
| | **DVM** | 0.7945 | **0.8102** | **0.6883** | 0.5699 |
| **Minkowski** | **RO** | 0.8806 | **0.8912** | 0.9188 | **0.9423** |
| | **LR** | 0.8806 | **0.8912** | **0.7182** | 0.6416 |
| | **KA** | 0.7904 | **0.8354** | 0.9188 | **0.9423** |
| | **DVM** | 0.7956 | **0.8397** | 0.6956 | **0.9235** |

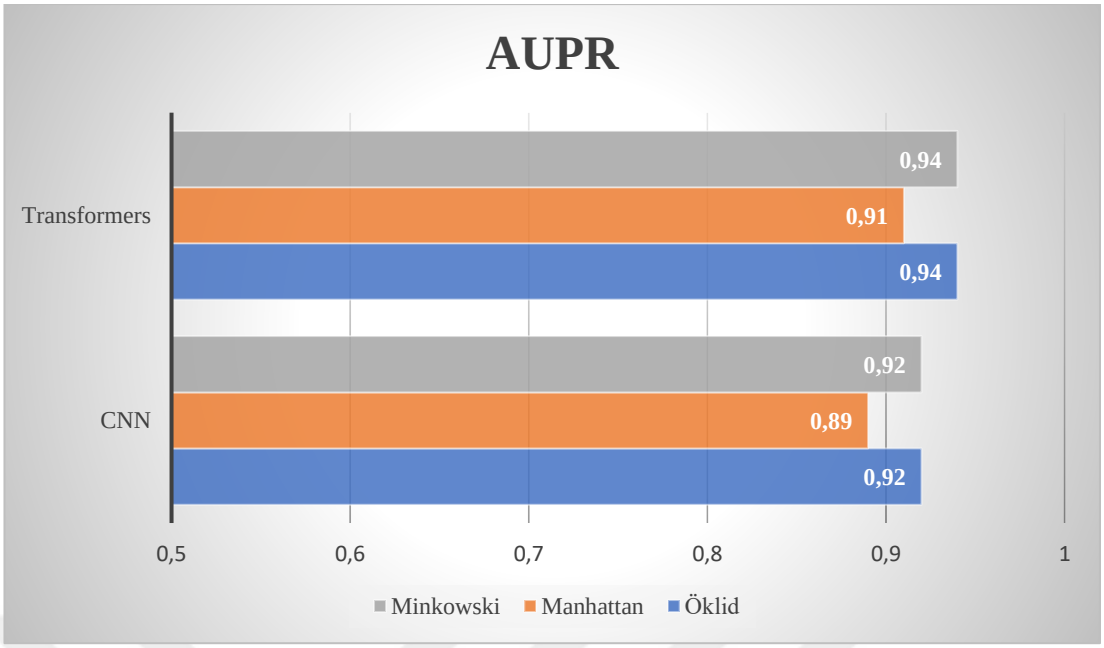
AUC ve AUPR, İHE tahmini alanında sıklıkla hesaplanan alanlardır. Çizelge 5.1. 'e göre iki ayrı veri seti değerlendirildiğinde RO sınıflandırma modeli ve transformers ağının kullanıldığı veri setinin sonuçlarındaki yüksek değerlerleri ön plana çıkmaktadır. Doğruluk, f1 skoru ve hassasiyet değerleriyle birlikte ele alındığında yine

en dengeli sonuçların RO modeline ait olduğu gözlemlenmektedir. Çizelge 5.1.' de veri setlerinden elde edilen AUC (Roc Eğrisi altında kalan alanın) değeri ve AUPR (precision-recall eğrsinin altında kalan alan) hesaplanmıştır. İHE tahminin en önemli değerlendirilmelerinden bir diğerleri de AUC ve AUPR sonuçlarıdır. Şekil 5.1 ve Şekil 5.2'de AUC ve AUPR değerlerinin karşılaştırılması görselleştirilerek desteklenmiştir.

Veriler, derin öğrenme ve makine öğrenme yöntemleri ile işlenmiş ve performans metrikleri karşılaştırılmıştır. Çalışmanın sonuçları için farklı performans metrik değerlerin hesaplanmasının ve karşılaştırılmasının sebebi farklı açılardan da sonuçların doğruluğunu ölçmek olmuştur. Çizelge 4.1., Çizelge 4.2. ve Çizelge 5.1. değerleri karşılaştırıldığında genel olarak Transformers ağı ile öznitelik çıkarımı sonuçlarının yüksek ve dengeli olduğu gözlemlenmektedir. Üç çizelgenin sonuçlarına göre en yüksek performans metriği, RO sınıflandırma modeline aittir. Dolayısıyla tüm sonuçlar karşılaştırıldığında sadece doğruluk veya sadece hassasiyet değerleri gibi tek değere odaklanılarak yapılan tartışmalar yanıltıcı olabilmektedir. Bu sebeple hassasiyet, kesinlik, F1 skoru değerleri ile AUC ve AUPR değerleri de göz önüne alındığında; ilaçların iki boyutlu yapısal formlarının transformers ağı ile özniteliklerinin çıkarıldığı veri setine ait tahmin sonuçlarında artışın olduğu gözlemlenmiştir.




Şekil 5.1. CNN ve Transformers veri setleri RO modelline ait AUC değerlerinin karşılaştırılması




Şekil 5.2. CNN ve Transformers ağı veri setleri RO modeline ait AUPR değerlerinin karşılaştırılması

Bugüne kadar yapılan çalışmalar yeni makine öğrenmesi, derin öğrenme ve yapay zeka algoritmalarının ihtiyacının yanı sıra doğru etkileşimlerin bulunması için ham verilerin farklı temsillerine de gereksinim duyulduğunu açıkça göstermiştir. Rifaioglu vd. (2020) araştırmasında ilaçlar için bu konuda sıkça kullanılan temsil olan moleküler parmakizinin dışına çıkılan ilaç moleküler yapı görüntülerinin kullanıldığı DeepScreen yaklaşımı ile bu konuya çok yenilikçi bir bakış açısı kazandırılmıştır. Yüksek AUPR ve AUC skorları ile yeni ilaç ve hedef temsillerinin ne kadar önemli olduğu gösterilmiştir. Bu bağlamda, çalışmamızda klasik ilaç temsili olan Moleküler Parmakizi ve Morgan/Circular Parmakizi yerine Python Rdkit kütüphanesi ile ilaçların moleküler yapıları elde edilip sonrasında iki boyutlu yapı görüntülerinin boyutları eşitlenerek CNN ile 300 boyutlu öznitelikleri kullanılmıştır. Ve ayrı bir veri seti olarak da ilaçların iki boyutlu moleküler yapıları transformers ağı kullanılarak özellikleri çıkarılmış ve sonuçları karşılaştırılmıştır. Karşılaştırılan sonuçların neticesinde ilaç görüntülerinin transformers ağı aracılığıyla çıkartılan özellikleri ile RO algoritmasının f1 skoru ve diğer performans metriklerinin daha yüksek olduğu gösterilmiştir. Moleküler/ Morgan parmak izi gibi ilaç temsillerinin kullanılmamasının en temel sebebi, belirtilen temsiller de önemli bilgilerin kaybolmasıdır. Bu minvalde, tez kapsamında ilaçlar için Transformers ağının kullanılarak diğer yöntemler ile karşılaştırılmasının avantajı, ilacın kimyasal yapısındaki özelliklerinin kaybolmadan

korunması ve evrişimli sinir ağlarına göre oldukça geniş bir bakış açısı sağlamalarıdır(Zhang vd., 2022). İlaçların temsilinde, CNN ve transformers ağının kullanılmasının sebebi ise ilaçların doğada var olan formlarına en yakın hali ile özniteliklerinin belirlenebilmesidir ki bu da sonuçların değerlerini olumlu yönde etkilemektedir.

İHE tahmini alanının faydasının en yüksek olacağı alan ilaç yeniden konumlandırma çalışmalarıdır. Bu çalışmalara katkı sağlayacak literatürde var olan İHE tahmini modelleri genellikle ilaçların SMILES dizileriyle yapılan çalışmalardır. İlaçların sadece SMILES dizileri ile çalışılmasının dezavantajı, ilaçların moleküler temsillerinin oluşturulamamasından dolayı bu yaklaşımların hastalıkların tedavisinde verimli sonuçlar elde edemeyecekleridir. GraphDTA, Deep-MGT ve DeepScreen makalelerinde ilaçların farklı temsillerinin olumlu sonuçları yeni temsillere gerekliliğin önemini açıklamaktadır. (Nguyen vd., 2021.; Rifaioglu vd., 2020; Zhang vd., 2022). Bu bağlamda, bu tez çalışmasının literatüre önemli katkılarından biri de ilaç temsili için kullandığımız ilaçların iki boyutlu moleküler yapılarının görüntülerinin transformers ağının kullanıldığı sınıflandırma modelinde yüksek sonuçlara ulaşmış olmasıdır.

Tez çalışmasını RO, DVM, LR gibi makine öğrenmesi tabanlı bir çalışma üzerine inşa edilmesinin sebebi ise İHE tahmini konusunda sıklıkla kullanılması ile birlikte iyi ve uygulanabilir sonuçların üretilmesidir. Diğer bir yandan, ilaç ve hedef verileri temsillerinde kullanılan karmaşık derin öğrenme yöntemleri sebebiyle veri setleri birleştirildikten sonra RO, DVM, algoritmaları derin öğrenme yerine tercih edilmiştir. Böylece protein ve ilaçların temsillerinden çıkardığımız özniteliklerin de tekrar performans sonuçlarını etkileyebilecek şekilde yanlış sonuçlar oluşturması da engellenmiş olmaktadır (Atas ve Doğan, 2022). Diğer yönden verilerin genelinde tahmin sonuçları yüksek olan RO kullanılmasının sebebi ise kullanımının kolaylığı ve karar ağacı sınıflandırıcısına dayalı bir topluluk öğrenmesi olmasıdır. Bu sebeple, yüksek kapsamlı özelliklerle, sonuçları daha doğru bir şekilde tahmin eder ve aşırı uyuma daha az eğilimlidir.

Tez çalışmamız için kullanılan yöntemler ve sonuçlar neticesinde çalışmamıza ek olarak ilaç yeniden konumlandırma ve moleküler docking çalışmaları ile

zenginleştirilerek yeni İHE çiftlerinin oluşturulması mümkündür. Yapılan çalışmalarda sonuçların yüksek çıkması durumunda ıslak laboratuvar uygulamaları ile desteklenerek bir sonraki aşamaya geçilebilir. Biyoinformatik ve keminformatik uygulamaların sonuçları in vitro çalışmalar ile desteklendiğinde anlam kazanacaktır. Bu sebeple, multidisipliner çalışmalara daha fazla ihtiyaç bulunmaktadır.

Yapılan çalışmada, ikili sınıflandırma sorunu olarak ele alınan İHE tahmini problemi ileri çalışmalar için tez çalışmasında kullanılan yöntemler ile birlikte çıkış veri setinin geliştirilmesi, doğru bağlanma afinite değerleri bulunması ve sonrasında da bağlanma yerlerinin kesinleştirilmesiyle doğru sonuçlara ulaşma olasılığı arttırılabilir.

Sonuç olarak, farklı yöntemler ile oluşturulmuş temsiller, farklı sınıflandırma stratejileri ile birlikte İHE tahmininin verimini arttırmanın etkili bir yol olduğunu gösterir. İHE tahmini metotları sürekli gelişecek ve geliştirilmeye açık olacaktır. Her geçen gün teknolojinin hızına yetişebilen yöntemler, hastalıklar için daha iyi tedavilerle sonuçlanan ilaçların geliştirilmesine katkı sağlayacaktır. İHE tahmininin doğru sonuçları ilaç sanayisinin maliyetlerinde düşüşe ve hastalarda iyileşme oranlarında olumlu etkiye sahip olacaktır. Bu bağlamda İHE tahmini üzerinde yapılacak çalışmaların geliştirilmesi önemli bir etkiye sahiptir. Makine öğrenmesi ve yapay zeka ile hızlandırılan çalışmaların geliştirilmesi sonucunda doğru ilaç hedef etkileşimleri, Alzheimer hastalığı tedavisinde kullanılan ilaçlar ile birlikte hastalık hastada keşfedildikten sonra sadece ilermesinin engellenmesi ile kalınmayacak, bu hastalığın gerileyebilmesi açısından da bir umut olacaktır. Alzheimer hastalığı verilerinin kullandıldığı ve iyi sonuçların elde edilen tez çalışmasının, tedavide kullanılmak üzere in vitro deneyler ile ilaç yeniden konumlandırılması alanına entegre olabilmesi amacıyla geliştirilmesine ve ileri çalışmalara ihtiyacı bulunmaktadır.

## KAYNAKLAR


Asgari, E., & Mofrad, M. R. K. (2015). Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLoS ONE*, *10*(11), 141287. https://doi.org/10.7910/DVN/JMFHTN

Ashburn, T. T., & Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*, *3*(8), 673-683.

Atas, H., & Doğan, T. (2022). How to Approach Machine Learning-based Prediction of Drug/Compound-Target Interactions. *bioRxiv*, 2022.05.01.490207. https://doi.org/10.1101/2022.05.01.490207

Ballard, C., Gauthier, S., Corbett, A., Brayne, C., Aarsland, D., & Jones, E. (t.y.). Alzheimer's disease. *www.thelancet.com*, *377*, 1019-1050. https://doi.org/10.1016/S0140

Ballard, C., Gauthier, S., Corbett, A., Brayne, C., Aarsland, D., & Jones, E. (2011). Alzheimer's disease. *the Lancet*, *377*(9770), 1019-1031.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*, 1-13.

Chu, Y., Kaushik, A. C., Wang, X., Wang, W., Zhang, Y., Shan, X., Salahub, D. R., Xiong, Y., & Wei, D. Q. (2021). DTI-CDF: A cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. *Briefings in Bioinformatics*, *22*(1), 451-462. https://doi.org/10.1093/bib/bbz152

Consortium, U. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*, *47*(D1), D506-D515.

Danielsson, P.-E. (1980). Euclidean distance mapping. *Computer Graphics and image processing*, *14*(3), 227-248.

Doytchinova, I. (2022). Drug Design—Past, Present, Future. Içinde *Molecules* (C. 27, Sayı 5). MDPI. https://doi.org/10.3390/molecules27051496

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*(5-6), 352-359. https://doi.org/10.1016/S1532-0464(03)00034-0

Druker, B. J. (2002). STI571 (Gleevec™) as a paradigm for cancer therapy. *Trends in molecular medicine*, *8*(4), S14-S18.

Du, B. X., Qin, Y., Jiang, Y. F., Xu, Y., Yiu, S. M., Yu, H., & Shi, J. Y. (2022). Compound–protein interaction prediction by deep learning: Databases, descriptors and models. Içinde *Drug Discovery Today* (C. 27, Sayı 5, ss. 1350-1366). Elsevier Ltd. https://doi.org/10.1016/j.drudis.2022.02.023

Ezzat, A., Wu, M., Li, X., & Kwoh, C. K. (2019). Computational Prediction of Drug-Target Interactions via Ensemble Learning. Içinde *Methods in Molecular Biology* (C. 1903, ss. 239-254). Humana Press Inc. https://doi.org/10.1007/978-1-4939-8955-3_14

Ezzat, A., Wu, M., Li, X. L., & Kwoh, C. K. (2017). Drug-target interaction prediction using ensemble learning and dimensionality reduction. *Methods*, *129*, 81-88. https://doi.org/10.1016/j.ymeth.2017.05.016

Ezzat, A., Wu, M., Li, X. L., & Kwoh, C. K. (2018). Computational prediction of drug-target interactions using chemogenomic approaches: An empirical survey. *Briefings in Bioinformatics*, *20*(4), 1337-1357. https://doi.org/10.1093/bib/bby002

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., & Cai, J. (2018). Recent advances in convolutional neural networks. *Pattern recognition*, *77*, 354-377.

Hasan Mahmud, S. M., Chen, W., Jahan, H., Dai, B., Din, S. U., & Dzisoo, A. M. (2020). DeepACTION: A deep learning-based method for predicting novel drug-target interactions. *Analytical Biochemistry*, *610*. https://doi.org/10.1016/j.ab.2020.113978

He, T., Heidemeyer, M., Ban, F., Cherkasov, A., & Ester, M. (2017). SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. *Journal of Cheminformatics*, *9*(1). https://doi.org/10.1186/s13321-017-0209-z

Hopkins, A. L. (2008). Network pharmacology: The next paradigm in drug discovery. Içinde *Nature Chemical Biology* (C. 4, Sayı 11, ss. 682-690). Nature Publishing Group. https://doi.org/10.1038/nchembio.118

Hu, P.-W., Chan, K. C. C., & You, Z.-H. (2016). Large-scale prediction of drug-target interactions from deep representations. *2016 international joint conference on neural networks (IJCNN)*, 1236-1243.

Huang, K., Fu, T., Glass, L. M., Zitnik, M., Xiao, C., & Sun, J. (2021). DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, *36*(22-23), 5545-5547. https://doi.org/10.1093/BIOINFORMATICS/BTAA1005

Huang, K., Xiao, C., Glass, L., & Sun, J. (2020). *MolTrans: Molecular Interaction Transformer for Drug Target Interaction Prediction*. https://doi.org/10.1093/bioinformatics/btaa880

Huang, Y., You, Z., & Chen, X. (2018). A Systematic Prediction of Drug-Target Interactions Using Molecular Fingerprints and Protein Sequences. *Current Protein & Peptide Science*, *19*(5), 468-478. https://doi.org/10.2174/1389203718666161122103057

Kanehisa, M. (2002). The KEGG database. *'In Silico'Simulation of Biological Processes: Novartis Foundation Symposium 247*, *247*, 91-103.

Kaushik, A. C., Mehmood, A., Dai, X., & Wei, D. Q. (2020). A comparative chemogenic analysis for predicting Drug-Target Pair via Machine Learning Approaches. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-63842-7

Kim, J., Park, S., Min, D., & Kim, W. (2021). Comprehensive survey of recent drug discovery using deep learning. *International Journal of Molecular Sciences*, *22*(18). https://doi.org/10.3390/ijms22189983

Lane, C. A., Hardy, J., & Schott, J. M. (2018). Alzheimer's disease. Içinde *European Journal of Neurology* (C. 25, Sayı 1, ss. 59-70). Blackwell Publishing Ltd. https://doi.org/10.1111/ene.13439

Lee, I., Keum, J., & Nam, H. (2019). DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Computational Biology*, *15*(6). https://doi.org/10.1371/journal.pcbi.1007129

Li, L., Wierbowski, S. D., & Yu, H. (2023). DeepERA: deep learning enables comprehensive identification of drug-target interactions via embedding of heterogeneous data. *bioRxiv*, 2021-2023.

Lin, X., Zhao, K., Xiao, T., Quan, Z., Wang, Z.-J., & Yu, P. S. (t.y.). *DeepGS: Deep Representation Learning of Graphs and Sequences for Drug-Target Binding Affinity Prediction*. https://www.uniprot.org/

Liu, G., Singha, M., Pu, L., Neupane, P., Feinstein, J., Wu, H. C., Ramanujam, J., & Brylinski, M. (2021). GraphDTI: A robust deep learning predictor of drug-target interactions from multiple heterogeneous data. *Journal of Cheminformatics*, *13*(1). https://doi.org/10.1186/s13321-021-00540-0

Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007). BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, *35*(suppl_1), D198-D201.

Masters, C. L., Bateman, R., Blennow, K., Rowe, C. C., Sperling, R. A., & Cummings, J. L. (2015). Alzheimer's disease. *Nature reviews disease primers*, *1*(1), 1-18.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. http://arxiv.org/abs/1301.3781

Nanor, E., Agbesi, V. K., Wu, W.-P., & Agyemang, B. (2020). FEATURIZATION OF DRUG COMPOUNDS AND TARGET PROTEINS FOR DRUG-TARGET INTERACTION PREDICTION. *International Journal of Scientific and Research Publications (IJSRP)*, *10*(2), p9813. https://doi.org/10.29322/ijsrp.10.02.2020.p9813

Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., & Venkatesh, S. (t.y.). *GraphDTA: predicting drug-target binding affinity with graph neural networks*. https://doi.org/10.5281/zenodo.3603523

O'Brien, R. J., & Wong, P. C. (2011). Amyloid precursor protein processing and alzheimer's disease. *Annual Review of Neuroscience*, *34*, 185-204. https://doi.org/10.1146/annurev-neuro-061010-113613

Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: Deep drug-target binding affinity prediction. *Bioinformatics*, *34*(17), i821-i829. https://doi.org/10.1093/bioinformatics/bty593

Pahikkala, T., Airola, A., Pietilä, S., Shakyawar, S., Szwajda, A., Tang, J., & Aittokallio, T. (2015). Toward more realistic drug-target interaction predictions. *Briefings in Bioinformatics*, *16*(2), 325-337. https://doi.org/10.1093/bib/bbu010

Park, Y., & Marcotte, E. M. (2012). Flaws in evaluation schemes for pair-input computational predictions. *Nature methods*, *9*(12), 1134-1136.

Peng, J., Li, J., & Shang, X. (2020). A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. *BMC Bioinformatics, 21*. https://doi.org/10.1186/s12859-020-03677-1

Podder, A., Pandit, M., & Narayanan, L. (2018). Drug target prioritization for Alzheimer's disease using protein interaction network analysis. *OMICS: A Journal of Integrative Biology*, *22*(10), 665-677.

Raghavendra, M., Rudolph Raj, J., & Seetharaman, A. (2012). A study of decrease in R&D spending in the pharmaceutical industry during post-recession. *International Journal of Academic Research Part B*, *4*(5), 2075-4124. https://doi.org/10.7813.2075-4124.2012/4-5/B.6

Rifaioglu, A. S., Nalbat, E., Atalay, V., Martin, M. J., Cetin-Atalay, R., & Doğan, T. (2020). DEEPScreen: high performance drug-target interaction prediction with convolutional neural networks using 2-D structural compound representations. *Chemical Science*, *11*(9), 2531-2557. https://doi.org/10.1039/c9sc03414e

Sachdev, K., & Gupta, M. K. (2019). A comprehensive review of feature based methods for drug target interaction prediction. Içinde *Journal of Biomedical Informatics* (C. 93). Academic Press Inc. https://doi.org/10.1016/j.jbi.2019.103159

Scheltens, P., Blennow, K., Breteler, M. M. B., de Strooper, B., Frisoni, G. B., Salloway, S., & Van der Flier, W. M. (2016). Alzheimer's disease. Içinde *The Lancet* (C. 388, Sayı 10043, ss. 505-517). Lancet Publishing Group. https://doi.org/10.1016/S0140-6736(15)01124-1

Shi, H., Liu, S., Chen, J., Li, X., Ma, Q., & Yu, B. (2019). Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics*, *111*(6), 1839-1852. https://doi.org/10.1016/j.ygeno.2018.12.007

Singh, A., Yadav, A., & Rana, A. (2013). K-means with Three different Distance Metrics. *International Journal of Computer Applications*, *67*(10).

Song, T., Zhang, X., Ding, M., Rodriguez-Paton, A., Wang, S., & Wang, G. (2022). DeepFusion: A deep learning based multi-scale feature fusion method for predicting drug-target interactions. *Methods*, *204*, 269-277. https://doi.org/10.1016/j.ymeth.2022.02.007

Taşcı, E., & Onan, A. (2016). K-en yakın komşu algoritması parametrelerinin sınıflandırma performansı üzerine etkisinin incelenmesi. *Akademik Bilişim*, *1*(1), 4-18.

Thafar, M. A., Olayan, R. S., Albaradei, S., Bajic, V. B., Gojobori, T., Essack, M., & Gao, X. (2021). DTi2Vec: Drug–target interaction prediction using network embedding and ensemble learning. *Journal of Cheminformatics*, *13*(1). https://doi.org/10.1186/s13321-021-00552-w

Thafar, M. A., Thafar, M. A., Olayan, R. S., Olayan, R. S., Ashoor, H., Ashoor, H., Albaradei, S., Albaradei, S., Bajic, V. B., Gao, X., Gojobori, T., & Essack, M. (2020). DTiGEMS+: Drug-target interaction prediction using graph embedding, graph mining, and similarity-based techniques. *Journal of Cheminformatics*, *12*(1). https://doi.org/10.1186/s13321-020-00447-2

Wei, B., & Gong, X. (t.y.). *DeepPLA: a novel deep learning-based model for protein-ligand binding affinity prediction*. https://doi.org/10.1101/2021.12.01.470868

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, *28*(1), 31-36.

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), 1401-1409. https://doi.org/10.1021/acs.jproteome.6b00618

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. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, *46*(D1), D1074-D1082.

Xu, L., Ru, X., & Song, R. (2021). Application of Machine Learning for Drug–Target Interaction Prediction. Içinde *Frontiers in Genetics* (C. 12). Frontiers Media S.A. https://doi.org/10.3389/fgene.2021.680117

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., & Kanehisa, M. (2008). Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, *24*(13). https://doi.org/10.1093/bioinformatics/btn162

Zhang, P., Wei, Z., Che, C., & Jin, B. (2022). DeepMGT-DTI: Transformer network incorporating multilayer graph information for Drug–Target interaction prediction. *Computers in Biology and Medicine*, *142*. https://doi.org/10.1016/j.compbiomed.2022.105214

Zhou, W., Wang, Y., Lu, A., & Zhang, G. (2016). Systems pharmacology in small molecular drug discovery. İçinde *International Journal of Molecular Sciences* (C. 17, Sayı 2). MDPI AG. https://doi.org/10.3390/ijms17020246

Zhou, Y., Zhang, Y., Lian, X., Li, F., Wang, C., Zhu, F., Qiu, Y., & Chen, Y. (2022). Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic Acids Research*, *50*(D1), D1398-D1407.