İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

İNSAN VE HIV-1 PROTEİNLERİ ARASINDAKİ ETKİLEŞİMLERİN RASTGELE ORMAN YÖNTEMİ VE BİRLİKTE ÖĞRENME YAKLAŞIMI İLE TAHMİN EDİLMESİ

YÜKSEK LİSANS TEZİ

İsmail BİLGEN

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

OCAK 2014

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

İNSAN VE HIV-1 PROTEİNLERİ ARASINDAKİ ETKİLEŞİMLERİN RASTGELE ORMAN YÖNTEMİ VE BİRLİKTE ÖĞRENME YAKLAŞIMI İLE TAHMİN EDİLMESİ

YÜKSEK LİSANS TEZİ

İsmail BİLGEN
(504101549)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Doç. Dr. Zehra ÇATALTEPE

OCAK 2014

İTÜ, Fen Bilimleri Enstitüsü'nün 504101549 numaralı Yüksek Lisans Öğrencisi **İsmail BİLGEN**, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"İNSAN VE HIV-1 PROTEİNLERİ ARASINDAKİ ETKİLEŞİMLERİN RASTGELE ORMAN YÖNTEMİ VE BİRLİKTE ÖĞRENME YAKLAŞIMI İLE TAHMİN EDİLMESİ"** başlıklı tezini aşağıdaki imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı :	**Doç. Dr. Zehra ÇATALTEPE** İstanbul Teknik Üniversitesi	.....
Jüri Üyeleri :	**Yrd. Doç. Dr. Ömer Sinan Saraç** İstanbul Teknik Üniversitesi	.....
	Yrd. Doç. Dr. Arzucan Özgür Boğaziçi Üniversitesi	.....

Teslim Tarihi: **16 Aralık 2013**
Savunma Tarihi: **24 Ocak 2014**

Aileme,

ÖNSÖZ

Çalışmalarım sırasında bilgi ve tecrübelerini benden esirgemeyerek bana yardımcı olan ve manevi desteğini eksik etmeyen değerli hocam ve danışmanım Sayın Doç. Dr. Zehra Çataltepe'ye, konuyu baştan aşağı birlikte mütalaa ettiğimiz ve fedakarane yardımlarını hiçbir zaman eksik etmeyen değerli hocalarım Yrd. Doç. Dr. Ömer Sinan Saraç'a ve Yrd. Doç. Dr. Arzucan Özgür'e teşekkürlerimi sunarım.

Bugüne kadar maddi ve manevi her konuda bana destek olan sevgili aileme teşekkür ederim.

Ocak 2014

İsmail BİLGEN
(Bilgisayar Mühendisi)

İÇİNDEKİLER

Sayfa

KISALTMALAR

AIDS	**:** Acquired Immunodeficiency Syndrome Edinilmiş Bağışıklık Eksikliği Sendromu
AUC	**:** Area under the Curve
ÇS	**:** Çapraz Sağlama
ED	**:** Eşik Değer
HIV	**:** Human Immunodeficiency Virus İnsan Bağışıklık Yetmezliği Virüsü
KP	**:** Kısmi Pozitif
MAP	**:** Mean Average Precision
N	**:** Negatif
ORT	**:** Ortalama
P	**:** Pozitif
PPE	**:** Protein-Protein Etkileşimi
PRBE	**:** Precision-Recall Break-even Point
RO	**:** Rastgele Orman
SD	**:** Standart Sapma
YSA	**:** Yapay Sinir Ağları

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

İNSAN VE HIV-1 PROTEİNLERİ ARASINDAKİ ETKİLEŞİMLERİN RASTGELE ORMAN YÖNTEMİ VE BİRLİKTE ÖĞRENME YAKLAŞIMI İLE TAHMİN EDİLMESİ

ÖZET

Protein-protein etkileşimi canlı organizmaların yaşamını devam ettirmesinde hayati önem taşır. Birçok hücresel fonksiyon proteinlerin etkileşmesi ile gerçekleşir. İnsan ve virüse ait proteinlerin etkileşmesi de viral enfeksiyon oluşmasında rol oynar. Bu nedenle etkileşen protein çiftlerinin bilinmesi hem insan biyolojisini hem de viral enfeksiyonları anlamak açısından önemlidir.

Bu çalışmada HIV-1 virüsüne ve insana ait proteinlerin etkileşip etkileşmediğini tahmin etmek için yapay öğrenme teknikleri kullanıldı. HIV-1 virüsüne ait 17 protein, insana ait proteinler ile 354841 olası etkileşim çifti oluşturmaktadır. Bu olası protein çiftlerinin, gerçek dünyadaki etkileşim oranının 100'de 1 olması beklenir. Bütün bu olası çiftlerin gerçekten etkileşip etkileşmediğini deneysel olarak test etmek zamansal ve finansal kısıtlardan dolayı mümkün değildir. Bu yüzden hesaba dayalı yöntemler araştırmacılara, arama uzayını daraltmada ve iyi adaylar önermede yardımcı olur.

Kullanılan veri kümesindeki örnekler, biri insana diğeri HIV virüsüne ait olmak üzere protein çiftlerinden oluşmaktadır. Her protein çifti 18 boyutlu bir vektör ile temsil edilmiştir. Protein çiftleri pozitif, negatif ve kısmi pozitif olarak sınıflandırılmıştır. Uzmanlar tarafından arasında etkileşim olduğu deneysel olarak onaylanmış protein çiftleri pozitif olarak sınıflandırılmıştır. Kısmi pozitif olarak sınıflandırılan protein çiftleri bazı anahtar kelimelere göre bilimsel literatürden elle çıkarılmıştır. Bu anahtar kelimeler iki grupta ele alınmıştır. Birinci grup anahtar kelimeler 'interacts with' (ile etkileşime geçer), 'binds' (bağlar) gibi etkileşimi göstermesi bakımından güçlüdür. İkinci grup anahtar kelimeler ise 'upregulate' (artarak düzenler) ve 'inhibits' (durdurur) gibi doğrudan etkileşimi göstermemesi bakımından daha zayıftır. Kısmi pozitif protein çiftleri negatife nazaran pozitife daha yakındır, ancak uzmanlar tarafından onaylanmadığı için pozitif sayılamaz. İki proteinin etkileşmediğini göstermek neredeyse imkansızdır. Dolayısıyla etkileşmeyen protein çiftlerinin geniş kümesi yoktur. Sınıflandırma işleminin yapılabilmesi için gerekli olan negatif örnekler, pozitif ve kısmi pozitiflerden arta kalan protein çiftlerinden örnekleme yöntemi ile alınır. Örnekleme etkileşime girmeyen protein çiftlerinin çoğunlukta olduğu varsayımına dayanarak yapılır. Negatif örneklerin bu yolla seçilmesi yaygın olarak kullanılan bir yöntemdir.

Gözetimli yapay öğrenme yöntemleri sınıflandırılmış veriye ihtiyaç duyar. Üzerinde fazlaca çalışılmış organizmalar haricinde birçok organizma için, başarılı bir sınıflandırıcı geliştirmeye yetecek miktarda protein-protein etkileşim verisi bulunmaz. Bu da ek bilgi kullanmayı gerekli kılar. Bu tezde kullanılan veri kümesinde, ek bilgi literatürden çıkarılan kısmi pozitif protein çiftleridir. Kısmi pozitif örnekler uzmanlar tarafından doğrulanmamış olduğu için, gürültü içermeye yatkındırlar.

Bu çalışmada kısmi pozitiflerin daha etkili kullanılabilmesi için çeşitli yaklaşımlar geliştirildi. *Sade* adı verilen ilk yaklaşımda kısmi pozitif veri yok sayıldı. Model, pozitif ve örneklenen negatif protein çiftleri kullanılarak geliştirildi. İkinci yaklaşımda bütün kısmi pozitif örnekler doğrudan pozitif kabul edilerek eğitim kümesine dâhil edildi. Test kümesi ise sadece uzmanlar tarafından onaylanan pozitiflerle örneklenen negatiflerden oluşturuldu. Üçüncü yaklaşımda kısmi pozitifler, beraber öğrenme yapısında şu şekilde kullanıldı. Pozitif ve örneklenen negatif protein çiftleri kullanılarak ilk model oluşturuldu. Bu model ile kısmi pozitif örnekler sınıflandırıldı. Bu sınıflandırma işleminin sonucunda yüksek değerde sınıflandırılan örnekler eğitim kümesine eklenerek model yeniden eğitildi. Bu işlem eğitim kümesine eklenecek örnek kalmayıncaya ya da eklenecek örnek sayısı önemsiz düzeye gelinceye kadar devam ettirildi.

Rastgele Orman yöntemi kullanılarak gerçekleştirilen deneylerin sonucuna göre, en iyi performans kısmi pozitiflerin kullanılmadığı yaklaşımda elde edildi. Öte yandan, bütün kısmi pozitiflerin doğru kabul edilerek eğitim kümesine dâhil edilmesi performansı olumsuz yönde etkiledi ve kesinliği düşürdü. Kısmi pozitiflerin birlikte öğrenme yapısında kullanılması, tamamın doğru kabul edildiği yaklaşıma göre daha iyi sonuç verdi. Ayrıca bu yaklaşım ile kısmi pozitif örneklerin kullanılmasından doğan kesinlik değerindeki düşüşün de önüne geçildi. Ancak performans öngörüldüğü biçimde arttırılamadı. Pozitif örneklerin niteliği bunun başlıca sebebi olarak yorumlandı. Pozitif örneklerin yarısı kullanılarak geliştirilen modelin performansı, tamamının kullanıldığı duruma göre pek farklılık göstermedi. Bu sonuç, pozitif örneklerin birbirine benzediği ve insan-HIV arasındaki etkileşim kümesinin tamamını temsil edecek şekilde yeterince kapsayıcı olmadıkları fikrini verdi.

PREDICTING HUMAN-HIV1 PROTEIN-PROTEIN INTERACTIONS USING RANDOM FORESTS IN A CO-TRAINING APPROACH

SUMMARY

Protein-protein interactions are very important for maintaining the life of an organism. Many biological functions are carried out with the interactions of proteins. Interactions between human and virus proteins play roles in viral infections. Therefore, identifying interacting pairs of proteins is critical to understand both human biology and viral infections.

In this study, we used machine learning methods to predict interactions between human and HIV-1 proteins. HIV genome encodes for 17 proteins (two of them are actually precursors of the envelope (*env gp160*) and gag (*gag pr55*)), resulting in 354841 possible HIV-human pairings. Actual physical interactions among these possible pairs are expected to be only 1 in about 100. Due to financial and time constraints it is not possible to experimentally verify whether each pair really interacts. Therefore, computational methods are indispensable to help researchers narrow down the search space and to suggest good candidates to test experimentally.

We approached this issue as a classification problem. We used machine learning methods to classify instances as interacting or non-interacting. Instances in the dataset are protein pairs, where one protein belongs to HIV-1 and the other to human. Each pair is represented by an 18 dimensional feature vector. These features can be grouped into three types:

- Features extracted by considering the properties of the proteins that are involved in the interaction individually.
- Features that represent information about the proteins as a pair.
- Features extracted from human interactome.

Protein pairs are labeled as positive, partial positive and negative. The instances with the positive label are verified by experts. There are only 158 such pairs. Partial positive protein pairs, on the other hand, were manually curated from the literature. Each pair is associated with a keyword which describes an evidence of the interaction between proteins. Pairs with keywords that are strong indicative of interaction such as ‘interacts with’ and ‘binds’ are named as group-1, and those with keywords that weakly suggest an interaction such as ‘upregulates’ and ‘inhibit’ are named as group-2. These pairs are more likely to be positive than negative. However, the interactions between them have not yet been verified by experts. There are 2129 protein pairs which are labeled as partial positive where 553 pairs belong to group-1 and 1575 pairs belong to group-2. We randomly sampled 16000 pairs from the remaining unlabeled data of 352328 protein pairs and used them as negative with the assumption that these

are highly enriched for non-interacting pairs. It is possible that some of them are interacting pairs, but evidence for their interaction has not been found yet.

We applied *Multi-layer Perceptron* and *Random Forest* machine learning techniques to predict interacting proteins. For training the machine learning models and calculating the performance, we used 5-fold cross-validation. We used *WEKA* and *R* software environments for implementation of the project.

Since the positive and negative classes are highly unbalanced in size, we applied sampling methods to reduce the difference between them. In *WEKA*, we used *SpreadSubSample* filter to balance classes. As a pre-process filter, it provides sampling of intended amount of instances from each class. In *R*, we used *sample* base method without replacement. Because the size of negative class is excessively larger than positive, we sampled only 16000 instances from it.

We investigated strategies for using partial positive instances efficiently. First strategy was called *naive* where the partial positive data is ignored. Training and testing was carried out by using only positive and sampled negative instances. In the second strategy, all partial positive data was included in the training set as positives. They were only used in training the model but not in testing. Test set consisted of positives validated by experts and sampled negative pairs. In the third strategy, we neither ignored the partial positives nor accepted them as positives. We applied the Random Forest algorithm in a co-training set-up as follows. We used positive data and sampled negative data to train an initial model. Then, we used this model to classify the partial positive instances and the ones that were predicted as positive with high confidence were added to the positive training set for the next iteration. This process was iterated several times until there were no more protein pairs to be added to the training set.

We evaluated results using *Mean Average Precision* (MAP), *Precision-recall Break-even Point* (PRBE) and *Area under the ROC Curve* (AUC) performance metrics. MAP provides a measure of quality and it is the mean of values of average precision at different recall levels. PRBE is the value(s) of cut-off(s) where precision and recall are equal. In other words, it is the value of points where precision-recall curve cuts the diagonal of the graph. PRBE can have multiple values since the precision-recall curve can intersect with the diagonal more than once. In that case, the largest PRBE value is considered. AUC is the area under the ROC curve. ROC (receiver operating characteristic) curve is obtained by plotting true positive rate as a function of false positive rate for different threshold values. It assesses the discriminative power of the model independent of the threshold. AUC gives a single value of averaged performance score for the ROC curve.

Supervised machine learning methods require labeled data to train the model. For most of the organisms except well-studied ones, there is no sufficient protein-protein interaction data to develop a successful classifier. Therefore, auxiliary information is essential. In the human-HIV protein-protein interaction dataset used in this thesis, the auxiliary information is partial positive protein pairs which are curated from the literature. Since the interaction between partial positive protein pairs have not yet been verified, they are prone to noise. As a result of our experiments using Random Forest classifier, the best performance is obtained by ignoring the partial positive instances (naive approach). Accepting all partial positive instances as true and using them in the training set decreased the performance in all performance metrics. However, using partial positive instances in a co-training set-up minimized

their negative effect on performance and stopped the decrease in precision either. We proposed to increase the performance of the model using partial information but it didn't match our expectations. We reduced the size of the positive training data by half and the performance was not affected. This suggests that the instances in positive set are similar to each other and are not comprehensive enough to represent the whole set of human-HIV interactions.

1. GİRİŞ

Proteinler canlıların yapıtaşını oluşturur. Hücrede DNA replikasyonu, kimyasal reaksiyonların katalizörlüğü (enzim), hücre sinyalleme ve ligand taşıma gibi birçok fonksiyonun yerine getirilmesinde görev alırlar. Bu görevleri yerine getirirken tek başına hareket etmez, başka protein ya da moleküllerle etkileşime girerler [1]. Proteinler belirli bir fonksiyonu yerine getirmek üzere başka proteinlerle bir araya gelerek büyük moleküler makineleri oluşturur [2]. Proteinlerin bu şekilde bir araya gelerek fiziksel bağlantı kurmasına protein-protein etkileşimi denir. Proteinler arasındaki etkileşmeyi çözmek biyolojik fonksiyonların altında yatan sebepleri anlamamıza yardımcı olur. Protein-protein etkileşimi belli bir organizmaya ait proteinler arasında olabileceği gibi, farklı iki orgnizmaya ait proteinler arasında da olabilir. İnsan ile enfekte olan virüs proteinleri arasında bu denli bir etkileşim söz konusudur. Virüs bulaşmak, hücreye girmek ve yeni nesil viryonlarını üretmek için konak hücreye ihtiyaç duyar [3]. Bu bakımdan, konak ve patojen proteinleri arasındaki etkileşimi çözmek hastalığın biyolojik yolunu çözmede, uygun ilaç ve tedavi yolları geliştirmede yardımcı olur.

Protein-protein etkileşimlerinin deneysel olarak bulunması çok zaman alıcı ve masraflı bir iştir. Bundan dolayı, hesaplamalı yöntemler ile etkileştiği tahmin edilen protein çiftleri, araştırmacılara deneylerine nereden başlamaları konusunda yardımcı olur. Bunun yanında, gözetimli yöntemler fazla sayıda veri kümesine ihtiyaç duyarlar. Üzerinden çok çalışılmış organizmalar dışında, çoğu organizma için yeterli miktarda güvenilir protein etkileşim verisi yoktur. Bu yüzden, yarı-gözetimli yöntemlere ihtiyaç duyulur [4].

Bu çalışmada, insan ile HIV-1 proteinleri arasındaki ilişkiyi tahmin etmede yapay öğrenme yöntemleri kullanıldı. Veri kümesi biri insana diğeri HIV-1 virüsüne ait protein çiftlerinden oluşmaktadır. Her protein çifti pozitif, kısmi pozitif ve negatif olarak sınıflandırılmıştır. Pozitif sınıfta, uzmanlar tarafından onaylanmıış az sayıda protein çifti bulunmaktadır. Kısmi pozitif sınıfta, bilimsel literatürde birlikte geçen

ancak hakkında yeterli deneysel kanıt bulunmayan protein çiftleri bulunmaktadır. Bu çalışmanın amacı veri kümesinde bulunan kısmi pozitif protein çiftlerini en etkili biçimde kullanacak stratejiyi belirleyerek, yapay öğrenme modelini geliştirmek ve doğru pozitif tahminlerin sayısını arttırmaktır.

Bu bölümde HIV virüsü ve protein-protein etkileşimleri ile ilgili daha detaylı bilgi verilecek ve tezin organizasyonundan bahsedilecektir.

1.1 HIV (Human Immunodeficiency Virus)

HIV (human immunodeficiency virus, insan bağışıklığı yetmezliği virüsü), AIDS'e (acquired immunodeficiency syndrome, edinilmiş bağışıklık eksikliği sendromu) sebebiyet veren bir virüstür. HIV, hayat döngüsünü devam ettirebilmek için konak insana ihtiyaç duyar. HIV virüsü insanlarda bağışıklık sisteminde zaafa yol açarak fırsatçı patojenlere kapı aralar.

Edinilmiş bağışıklık eksikliği sendromu (AIDS ya da EBES) etkeni olan İnsan Bağışıklık Yetmezliği Virüsü (HIV) ilk kez 1981 yılında keşfedilmiştir. HIV-1 ve HIV-2 olmak üzere iki major tipi vardır. HIV-2, HIV-1'e göre daha az patojen olup daha sınırlı bir coğrafyada etki gösterir.

HIV-1 tek sarmal RNA genomuna sahiptir. Sadece 15 proteini kodlar. Bu yüzden konak insan hücreye ihtiyaç duyar [5].

HIV-1 ve HIV-2 kan, semen ve vajinal sıvılar ile birlikte anneden bebeğine doğum sırasında ya da emzirme sırasında bulaşabilir. Enfeksiyonu üç fazda gerçekleşir [6]:

- Geçici akut retroviral sendrom,
- Klinik latent dönem,
- AIDS gelişimi.

HIV enfeksiyonunun ardından kanda HIV'e özgü virolojik ve immünolojik parametrelerdeki değişimler şu sıra ile gözlemlenir [6]:

- HIV RNA,
- HIV p24 antijen,

- HIV antikorları.

Bu üç göstergenin kanda saptanma zamanı değişiklik gösterebilir. Enfeksiyon gerçekleştikten sonra viral replikasyon olmasına rağmen HIV RNA, antijen ve antikor gözlenemez. 1-4 hafta arasında antijenler gözlemlenebilir düzeye ulaşır. Ancak HIV antikorları ancak 1-2 ay içerisinde tespit edilebilir düzeye ulaşır. HIV antikorlarının tespit edilemediği bu döneme pencere dönemi denir. Hastalığın tanısında önerilen testlerden biri HIV antikor testidir. Bir diğer test ise p24 antijen testidir. HIV antikor üretimini tetikleyen p24 antijeni HIV tarafından üretilir. Son zamanlarda antikor testleri ile birlikte p24 antijen testi de uygulanarak antikor oluşumunun başlamadığı pencere döneminde erken teşhis imkanı sağlanır [6].

HIV PCR yöntemi ile HIV virüsünün genetik oluşumları test edilir. Bu yöntem HIV RNA ve HIV DNA olmak üzere ikiye ayrılır. Kan ve organ nakli yapacak olan vericilere erken tanı imkanı vermesi nedeniyle HIV RNA uygulanır. HIV pozitif annelerden doğan bebeklere ise HIV DNA uygulanır [6].

HIV virüsü bulaştıktan sonra şiddetli belirtiler hemen gözlenmez. Asemptomatik adı verilen bu dönemde virüs yardımcı T hücreleri, makrofajlar ve dentritik hücreler gibi bağışıklık sistemi hücrerelerine, enfekte olur ve çoğalır. Özellikle yardımcı T hücrelerinden olan CD4+ T hücreleri bundan olumsuz etkilenirler [7, 29].

AIDS araştırmalarında en tartışmalı konulardan biri HIV enfeksiyonunda T hücrelerinin ölümüne sebep olan mekanizmayla alakalıdır. T hücrelerinin ölüm sebebi virüs sebebiyle doğal yıkıma uğraması olabilir. Mohri ve arkadaşlarına göre, HIV virüsü girdiği hücrede yüksek aktivasyon ve devire neden olur ve T hücrelerinin tükenmesi üretim düşüklüğünden ziyade hücre devrinin artmasından kaynaklanmaktadır [8]. Bir diğer ölüm sebebi olarak HIV-1 bulaşan insanlarda T hücrelerinin apoptoza (apoptozis *eng.*) uğraması öne sürülür. HIV-1 bulaşmış bünyede, virüsün girdiği ve girmediği hücrelerde apoptoza uğrama miktarı artar [9]. Apoptoz, programlanmış hücre ölümünün ana tiplerinden biridir. Vücutta ihtiyaç duyulmayan ve anormalleşmiş hücrelerden kurtulmanın normal yoludur. Gelişen bir embriyoda parmakların birbirinden ayrılması için parmak arasındaki hücreler apoptoz başlatırlar [30].

Normal insanlarda CD4+ T hücre sayısı 1000 hücre/μL'den fazla iken HIV bulaşmış kişilerde bu sayı 200 hücre/μL'ün altına düşer. Bu da bağışıklık sisteminde zaafa sebep olarak fırsatçı enfeksiyonlara kapı aralar [7].

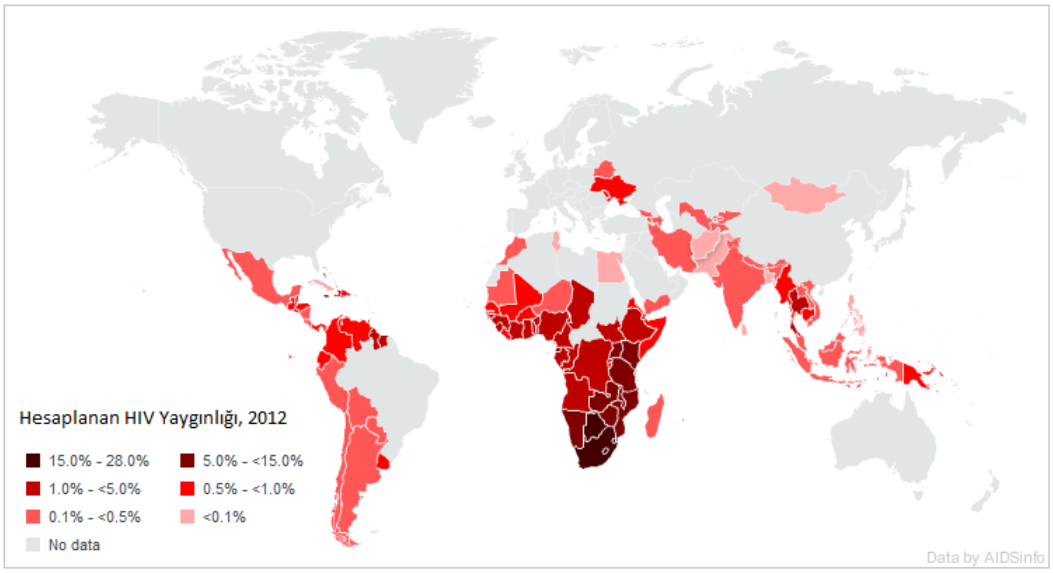
HIV'nin en sık bulaşma yolu cinsel temas olduğu için, HIV'den korunmanın en temel yolu, hastalığın enfekte olmadığı tek eşliliğe dayalı bir ilişki sürdürmektir. Hastalığın kan yoluyla da bulaşabildiği göz önününe alınacak olursa; tıraş bıçağı, diş fırçası gibi kişisel hijyen araçlarının paylaşılmaması önemlidir [7].

Yakın zamanda kullanılan rutin tedavi yöntemleri viral ters transkriptazı (reverse transcriptase *eng.*) ve proteaz enzimlerini (protease enzyme *eng.*) indirgerler. Virüse karşı kullanılan (antiretroviral *eng.*) ilaçlar virüsü baskılamaya yarasa da hastalığı yok etmeye yaramaz. Toksik oluşumu, metabolizmada düzensizlik ve HIV virüsünün ilaca karşı direnç kazanması gibi birçok sebepten ötürü alternatif tedavi yöntemlerine ihtiyaç duyulmaktadır. Çalışmalar daha çok virüs ile konak hücre arasındaki etkileşimleri tanımlamak noktasında yoğunlaşır. Çünkü diğer bütün virüsler gibi HIV de işlevlerini yerine getirebilmek için konak hücre ile etkileşime geçmek zorundadır [3].

UNIAIDS 2013 raporuna göre HIV ile yaşayan insanların sayısında sürekli bir artış olurken, HIV'e yeni yakalananların sayısında 1990'un sonlarından itibaren azalma olmuştur. HIV ile yaşayan insanların sayısındaki artış antiretroviral tedavi alanların sayısındaki artış ile doğru orantılı olabilir. WHO raporuna (2012) göre 2012 aralık ayında tespit edilen 9.7 milyon insan orta ve düşük gelirli ülkelerde antiretroviral tedavi gördü [10]. Bu rakam 2011 yılına göre 1.6 milyon artış olduğunu göstermektedir.

Dünya çapında HIV yaygınlığı UNIAIDS 2012 verilerine göre Şekil 1.1'deki gibidir.

HIV en yaygın olarak Sahra-altı Afrika ülkelerinde görülür. 2001 yılında HIV ile yaşayan insanların sayısı 20.3 milyon civarında iken 2009'da 22.5 milyona ulaşmıştır. Ancak HIV'e yeni yakalananların sayısında 2.2 milyondan 1.8 milyona düşüş vardır. Güney Asya (Hindistan, Pakistan, Bangladeş, Sri Lanka v.b.) ve Güneydoğu Asya (Asya kıtasıyla Okyanusya arasında bulunan ülkeler Brunei, Doğu Timor, Endonezya, Filipinler v.b.) ülkelerinde HIV ile yaşayan insanların sayısı 3.8 milyondan 4.1 milyona yükselmiştir. Yeni vakalarda ise önemli miktarda azalma vardır. Orta ve Güney Amerika'da HIV ile yaşayan insanların sayısı 1.1 milyondan 1.4 milyona çıkmıştır. Yeni vakalarda az bir düşüş vardır. Doğu Avrupa ve Orta Asya'da 760,000



Şekil 1.1: Dünya çapında HIV yaygınlığı [28].

civarı olan HIV ile yaşayan insanların sayısı 1.4 milyona çıkmıştır. Yeni vakalarda ise önemli bir düşüş vardır. Kuzey Amerika'da HIV ile yaşayan insanların sayısı 1.2 milyondan 1.5 milyona çıkmıştır. Yeni vakalarda ise az oranda artış vardır. Kuzey Afrika, Orta Doğu, Doğu Asya, Okyanusya, Karayipler ile Batı ve Orta Avrupa ülkeleri HIV ile yaşanlar insanların sayısı nispeten daha düşüktür. Verilen sayılar 2001 yılı ile 2009 yılının karşılaştırmasıdır.

Dünya genelinde HIV ile yaşayan insanların sayısı 2001 yılında 30.0 milyon civarında iken, 2012 yılında 35.3 milyon civarındadır. Türkiye'de ise HIV ile yaşayan insanların sayısı 2012 verilerine göre 3,900 - 8,000 civarında, yaygınlığı ise % 0.1'in altındadır [11].

1.2 Protein-Protein Etkileşimi

Canlı organizmaların yaşamlarını ve nesillerini devam ettirebilmeleri için birçok biyolojik fonksiyonu gerçekleştirmeleri gerekir. Proteinler bu fonksiyonların yerine getirilmesinde büyük bir role sahiptir. Proteinler çoğunlukla bir araya gelerek karmaşık, dinamik ve fizikokimyasal bağlantılara sahip moleküler makineleri oluşturur ve biyolojik fonksiyonları üstelenirler. Bu kompleks moleküler ilişkiyi çözmek protein protein etkileşimini anlamaktan geçer [2].

İnsan ve virüse ait etkileşen protein çiftlerinin bilinmesi enfeksiyonun nasıl oluştuğunu anlamaya ve buna bağlı olarak da yeni tedavi yöntemleri geliştirmeye yardımcı olur [12].

Protein-protein etkileşiminin tahmin edilmesi amacıyla deneysel yöntemler kullanılabilir. Ancak bu çok uzun süreli ve maliyetli bir işlemdir. Bazı yüksek kapasiteli yöntemler ile bir kerede çok sayıda etkileşim belirlemek mümkün olsa da sonuç veri kümeleri çoğunlukla eksiktir ve yüksek yanlış pozitif ve yanlış negatif ihtiva eder [4]. Bu nedenle hesaplamalı yöntemlere ihtiyaç duyulur. Hesaplamalı yöntemler ile çok sayıda protein çifti arasından fiziksel etkileşimde olma olasılığı en yüksek olanlar tahmin edilerek, deneysel yöntemlerde kullanılacak test sınıfı sınırlandırılabilir, ya da öncelik verilmesi gereken protein çiftleri oluşturulabilir.

Protein-protein etkileşiminin tahmin edilmesi ikili sınıflandırma problemi teşkil eder. İki protein arasında etkileşimin olması pozitif, olmaması negatif olarak nitelendirilir. İki proteinin etkileştiği deneysel olarak bulunabilir. Ancak denenmesi gereken çok sayıda protein çiftinin olması, deneylerin de maliyetli ve zaman alıcı olmasından dolayı pozitif örneklerin sayısı genelde azdır. İki proteinin etkileşmediğini gösteren raporlara ise pek rastlanmaz. Bundan dolayı yapay öğrenmede kullanmak üzere negatif etkileşim kümesi oluşturmak için çeşitli yöntemler geliştirilmiştir. Bunlardan biri negatif protein çiftlerinin, pozitif olduğu bilinen protein çiftleri dışında kalan örneklerden rastgele seçilmesidir. Negatif örneklerin bu şekilde seçilmesi gözetimli (supervised *eng.*) metotlarda eğitim kümesi oluşturulurken sıkça başvurulan bir yöntemdir [13]. Gerçekte her yüz protein çiftinden sadece birinin etkileşimli olduğu tahmin edilir. Bu oran negatif protein çiftlerini seçerken kullanılabilir.

1.3 Tezin Organizasyonu

Giriş bölümünde protein-protein ilişkileri ile HIV virüsüne değinildi. HIV virüsünün yapısından, nasıl bulaştığından, etkilerinden, korunma yollarından, teşhis ve tedavi yöntemlerinden, dünya çapındaki yaygınlığından bahsedildi. Protein-protein ilişkilerinin hücresel fonksiyonların yerine getirilmesindeki önemine değinildi. Aynı organizmaya ait proteinlerin etkileşebileceği gibi farklı organizmalardan proteinlerin de etkileşebileceği anlatıldı. Ayrıca, protein-protein etkileşiminin virüs

bulaşmasındaki önemi ve özelde HIV-1 virüsünün bulaşmasındaki yeri belirtildi. Son olarak, bu çalışmanın alt yapısından ve amaçlarından bahsedildi.

İkinci bölüm literatür taramasına ayrıldı. Bu konuda yapılmış çeşitli çalışmalardan bahsedildi.

Üçüncü bölümde bu çalışmada kullanılan yöntemlere değinildi. Kullanılan veri kümesi ve veri kümesindeki özniteliklerin anlamı anlatıldı. Ayrıca, kullanılan yapay öğrenme yöntemleri hakkında ayrıntılı olarak bilgi verildi.

Dördüncü bölümde uygulanan deneyler ve sonuçları verildi. Bu deneylerde kullanılan değerlendirme ölçütlerinden bahsedildi.

Son olarak beşinci bölümde sonuç ve öneriler sunuldu.

2. LİTERATÜR TARAMASI

HIV-1 ve konak insanın hücresel proteinleri arasındaki etkileşimin geniş çaplı kümesini tahmin etmeye yönelik ilk teşebbüs Tastan ve arkadaşları [12] tarafından ortaya konulur. Bu çalışmada çeşitli veri kaynakları kullanılarak gözetimli öğrenme mimarisi (supervised learning framework) uygulanır. Sınıflandırma yöntemi olarak RO (rastgele orman) kullanılır. Bilimsel literatürde bulunan insan ve HIV-1'e ait etkileşen proteinler NIAID veri bankasından alınır. Veri bankasında 1406 insan protein barındıran 2512 protein çifti bulunur. Bu protein çiftleri doğrudan fiziksel etkileşimi gösterme derecesine göre iki gruba ayrılır. Etkileşimi göstermesi bakımından daha güçlü anahtar kelimelerle ("interacts with" gibi) bahsi geçen protein çiftleri grup-1'de, diğerleri grup-2'de yer alır. Negatif örnekler, etkileşmediği kesin olarak bilinen protein çiftlerini içeren mevcut bir veri kümesi olmadığı için, etkileştiği bilenen örnekler dışındakilerden rastgele seçilir. Negatif küme oluşturulurken 1/100 oranı baz alınır. Veri kümesindeki örnekler için 35 özellik belirlenir [14]. 3-CV (cross validation) ile 10 kez tekrarlanan deney sonucunda ortalama MAP (Mean Average Precision) skoru 0.23 bulunur. AUC (Area Under the Curve) skoru da 0.9150 bulunur.

Yanjun ve arkadaşları [4] çalışmasında, bilinen doğrudan etkileşimli protein çiftlerinin az olduğunu ama bunun yanında aralarında etkileşim olduğu öngörülüp de hakkında yeterli deneysel kanıt bulunmayan protein çiftlerinin kayda değer miktarda olduğunu bildirir. Kısmi sınıflandırılmış (partially labeled) diye isimlendirilen bu veriyi çözüme katmak için yarı-gözetimli çoklu-görev (semi-supervised multi-task) mimarisi önerilir. Yarı-gözetimli yöntem yardımcı görevler olarak 3 farklı strateji ile uygulanır. Sınıflandırılmış eğitim verisi kullanılarak çok katmanlı algılayıcı ağı oluşturulur. Yardımcı görevler bu ağın katmanlarını paylaşır. Veri kümesi olarak önceki çalışmalarında açıklanan veri kümesini [12] kullanırlar. Yalnız önceki çalışmalarında veri kümesi elemanları 35 özellikle tanımlanırken, bu çalışmalarında 17 özellik çıkarılarak yalnız 18 özellik kullanılır. Önceki çalışmada literatürden elle çıkarılan 2512 protein çifti pozitif veri kümesini oluşturur. Bu çalışmada, bu pozitif veri

kümesinden 158 tanesi uzmanlara deneysel olarak onaylatılarak altın-standartta (gold standard) pozitif veri kümesi oluşturulur. Geri kalan ise kısmi-pozitif olarak anlatıldığı şekilde yardımcı görevlerle gözetimli sınıflandırıcının performansını arttırmak için kullanılır.

Aktif öğrenme, etkileşime giren protein çiftlerinin tahmin edilmesinde kullanılan yöntemlerden biridir. [15] çalışmasında, rastgele orman (random forest) metotu ile eğitilecek protein çiftlerinin seçilmesinde dört farklı strateji ile aktif öğrenme kullanılır. Çalışmada, tahmin edilen etkileşimlerin f-skoru (kesinlik (precision) ve anma (recall) değerlerinin harmonik ortalaması) aktif öğrenme kullanıldığı durumda, verinin rastgele seçilmesi durumuna göre %15 daha fazla bulunur.

Sistem biyolojinin çalışma alanlarından biri de biyolojik nesneler arasında tamamlanmış ağ yapısı oluşturmaktır. Ağ, düğümlerden ve aralarındaki ayrıtlardan oluşur. Düğümler proteinler ya da genler gibi biyolojik nesnelere denk gelir. Buna karşılık ayrıtlar protein etkileşimi ağında etkileşimi, gen düzenleyici ağında (gene regulatory network) düzenleyici protein ile düzenlediği gen arasındaki bağlantıyı, genetik ağda ise genetik ilişkiyi gösterir. Ağ yapısı bize bioyolojik fonksiyonların nasıl yürüdüğü ile ilgili önemli bilgi verir. Ağ oluştumayı amaçlayan birçok hesaplamalı yöntemin başarısı yüksek güvenirlikli verilerin az olması sebebiyle kısıtlıdır. [16] çalışmasında eğitim kümesini genişletmeyi amaçlayan iki yöntem geliştirilmiştir. Geliştirilen iki yöntem de yarı-gözetimli öğrenmeye dayalı, eğitim kümesinde sınırlı sayıda bulunan altın-standarttaki örnekleri, özenle seçilmiş ve yüksek güvenilirlikli yardımcı veriler ile arttırmayı hedefler. Birinci yöntem, tahmin yayılımı (prediction propagation), ile yerel modelin yüksek güvenirlikli tahminlerini bir başkasına yardımcı örnek olarak verir. Kavram olarak birlikte öğrenme (co-training) yöntemine benzer. İkinci yöntem, kernel başlatma (kernel initialization), birbirine en çok ve en az benzeyen nesnelere pozitif ve negatif eğitim kümesi elemanı olarak davranır. Bu yöntemlerle, mayaya ait birtakım protein-protein etkileşim ağları üzerindeki tahminlerde diğer temel yerel modellemelere göre önemli derece iyileşme gösterir. Diğer yöntemler tarafından düşük puanlanan bazı etkileşimleri de doğru sınıflandırmayı başarır. [16]

Mikrodizi tabanlı gen ifadesi belirleme (gene expression profiling) değişik kanserlerin tiplerinin sonuçlarını, prognoz ve belirli tedavilere karşı hassasiyeti tahmin etmede kullanılabilir [17, 18]. Ancak klinik verilerle desteklenen sınıflandırılmış örneklerin az

oluşu, protein etkleşim verisinde olduğu gibi gözetimli yöntemlerin etkili çalışmasını engeller. Shi ve arkadaşlarının çalışmasında [17], yarı-gözetimli LDS (low density separation *eng.*, düşük yoğunluk ayrımı *tr.*) [19] yöntemi kullanılarak kolon kanseri hastalarında kötüye gitme riski tahmin edilir. Çalışmanın sonucuna göre en gelişkin gözetimli SVM (support vector machine *eng.*, destek vektör makinesi) yöntemine göre yarı gözetimli LDS yöntemi, sınıflandırılmamış veriyi de kullanarak tahmin kesinliğini arttırır ve daha iyi performans sağlar.

3. YÖNTEM

Bu bölümde insan ve HIV-1 protein-protein etkileşim veri kümesi ile ilgili detaylı bilgi verilecek. Ayrıca, kullanılan yapay öğrenme yöntemleri ile beraber öğrenme yaklaşımı anlatılacak.

3.1 HIV-1 Human PPE Veri Kümesi

Bu çalışmada protein-protein etkileşimi veri kümesi olarak, Qi ve arkadaşları tarafından 2010 yılındaki çalışmada [4] kullanılan ve ilave web adresinde [20] sunulan veri kümesi kullanıldı. Bu veri kümesi aslında Taştan ile birlikte 2009 yılındaki çalışmalarında [12] insan ve HIV-1 protein-protein etkileşiminin evrensel kümesini çıkarılmasına yönelik hazırlanır ve ikili sınıf problemi haline getirilir. Protein çiftleri NIAID [21] veri tabanından alınır. NIAID veritabanında bilimsel literatürden elle çıkarılmış HIV-1 ve insana ait etkileşen protein çiftleri bulunur. Veritabanında 1406'sı insana ve 17'si HIV-1'e ait, 2620 protein çifti bulunur (17 Kasım 2007'deki güncellemeden önce). Bu protein çiftleri ilişkili olduğu anahtar kelimeye göre iki gruba ayrılır. Birinci gruptakiler, ikinci gruba göre pozitif etkileşimli olmaya daha yakın anahtar kelimeler içerir. Örneğin "etkileşime girer" (interacts with), "bağlar" (binds) gibi. Grupları oluştururken kullanılan kelimenin tam listesi ilave dökümanda bulunabilir [14]. 2010'daki çalışmada üzerinde bazı değişiklikler ve geliştirmeler yapılır. Öncelikli olarak, literatürden elle çıkarılan protein çiftlerinden bir kısmı uzmanlara gönderilir ve deneysel olarak onaylanmış gold-standart protein etkileşim çiftleri oluşturulur. Ayrıca 35 olan özellik sayısı 18'e düşürülür.

Veri kümesindeki elemanlar biri insansa, diğeri HIV-1 virüsüne ait olan protein çiftlerinden oluşur. Bu protein çiftleri "etkileşir" ya da "etkileşmez" şeklinde sınıflandırılır. Her bir protein çifti 18 özellik ile temsil edilir. Bu özelliklerin bir kısmı yalnız HIV-1 veya insan proteinini, bir kısmı da ikisi arasındaki ilişkiyi ilgilendirir. Özellikler kümesi bunları kapsar:

- Doku benzerliği. Eğer bir protein HIV-1 virüsüne duyarlı dokulardan birinde ifade edilmiyorsa HIV-1 proteinleri ile o protein arasında etkileşim olma ihtimali düşüktür.

- Topolojik özellik. İnsan proteinlerinin insan interaktomundaki topolojik özelliklerini tanımlar. Proteinlerin düğümü, etkileşimlerin de kenarı temsil ettiği yönsüz ağda derece (degree), kümelenme sabiti (clustering coefficient) ve aradalık merkeziliği (betweenness centrality) özelliklerinden yaralanılır.

- HIV-1 protein tipi.

- Dizi benzerliği (sequence similarity). HIV-1 ile insan proteinin (ya da etkileşimli olduğu komşu insan proteinin) arasındaki benzerliği tanımlar. Benzer yapıdaki proteinlerin etkileşme ihtimali daha yüksek olabilir.

- Posttranslasyonel modifikasyon benzerliği. Bazı protein etkileşimleri proteinlerin aynı posttranslasyonel modifikasyon durumunda olmasını gerektirir. Bu sebeple benzerlik ilişkisi ikili olarak kurulur. HIV-1 ile insan proteininin en az bir komşusunun aynı posttranslasyonel modifikasyon durumuna sahip olma durumu 1 olarak nitelendirilir.

- Gen ifadesi değişimi. HIV-1 bulaşmış ve bulaşmamış numunelerde, genlerin ortalama ifade seviyelerindeki değişim olarak ölçülür. Gen ifade seviyesi HIV-1 bulaşmış örnekte D^+, bulaşmamışta D^- olmak üzere değişim, $log(D^+/D^-)$ şeklinde hesaplanır.

- ELM-ligand özelliği. Ökaryotik lineer motif (ELM) veriyapısından [22] alınan kısa fonksiyonel dizi motiflerinden protein alanına ya da belirli bir protein sınıfına bağlanmaya aracı olan motifler çıkarılır. Eğer HIV-1 proteinlerinin dizisinde bir ELM motifi varsa ve ligand alanı karşılık gelen insan ortağında bulunuyorsa ya da insan ortağı o ligand sınıfından ise bu özellik 0 ile 1 arasında değer alır.

- Gen ontolojisi benzerliği. İki protein belirtim (annotation) kümesi arasındaki benzerlik şu şekilde bulunur. Birinci protein belirtim kümesindeki her bir terimin ikinci protein belirtim kümesindeki terimlerle benzerliklerin en yüksek olanı alınır. Her bir terim için bulunan en yüksek benzerlik değerlerinin ortalaması alınır. GO terimleri arasındaki benzerlik semantik benzerlik yöntemi, G-SESAME [23], ile

hesaplanır. Gen ontoloji çizgesinde terimlerin yalnız ortak atalarına bakılmaz, konumları ve bağlantı tipleri de baz alınır.

Veriyi sağlayan makalede [4] veri kümesi ile alakalı verilen sayılarla, veri üzerinde yaptığımız incelemeler sonucu elde ettiğimiz sayılarda farklılıklar gözlendi. Bunun üzerine makale yazarları ile kurulan iletişim sonucunda, bunun versiyon karışıklığından olduğu öğrenildi. Bu ilave web adresinde sunulan veri kümesindeki versiyon hatasının giderilmesine vesile oldu. Bu çalışmada veri kümesi ile ilgili verilen sayılar, makaledekinden farklılık arzetse de, veri kümesinin en güncel haline aittir.

HIV-1 virüsü 15 protein kodlar. Bunlara envelope (env gp160) and gag (gag pr55) proteinlerinin prekursörleri de eklenmiştir. 20873 insan proteini ile 354841 protein etkileşim çifti oluşturur. Bu veri kümesindeki eleman sayısı manasına gelir.

Veri kümesinde 384 adet uzmanlar tarafından test edilen protein çifti bulunur. Bunların 158 tanesi pozitif (etkileşim var) olarak sınıflandırılmıştır. 226 tanesi hakkında direk fiziksel etkileşim olduğu kanıtlanamamış dolaylı ya da şüpheli olarak nitelendirilmiştir. 384 adet uzman onaylı protein çiftinden 294 adedi grup 1'e, 87 adedi grup 2'ye aittir. 3 adedi iki grupta da bulunmaz. 847 adet protein çifti grup 1'in ve 1663 adet protein çifti de grup 2'nin içinde olmak üzere toplam 2512 adet literatürden elle çıkarılmış protein çifti bulunur. Uzmanlar tarafından onaylanan protein çiftleri çıkarıldığında 2129 adet protein çifti kalır $((847-294)+(1663-87))$. Bunlar kısmi pozitif protein çiftlerini oluşturur.

Özetlenecek olursa veri kümesinde bulunan protein çiftlerinin sayı değerleri çizelge 3.1'deki gibidir. Geriye kalan protein çiftlerinin sayısı, şüpheli olan 226 protein çiftinin de çıkarılması ile hesaplanmıştır.

Çizelge 3.1: HIV-insan protein-protein etkileşimi veri kümesi.

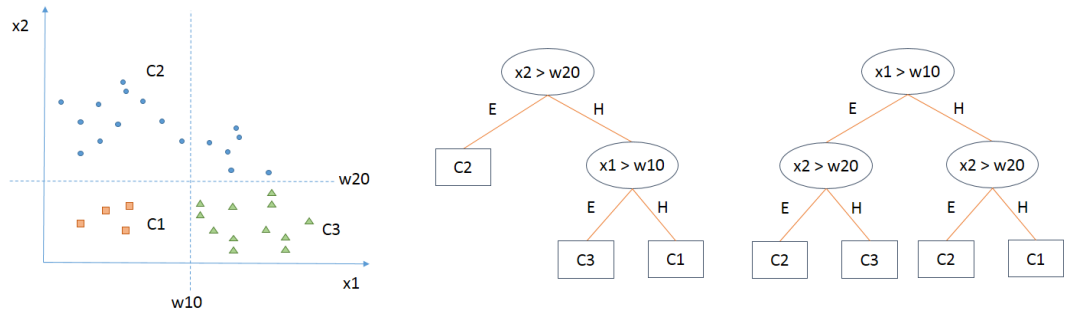
HIV-1 Proteinleri	İnsan Proteinleri	Pozitif (Altın Standart) Protein Çiftleri	Kısmi Pozitif Protein Çiftleri	Geriye Kalan Protein Çiftleri
17	20873	158	2129	352328

3.2 Yapay Öğrenme Yöntemleri

3.2.1 Karar Ağaçları

Karar ağaçları gözetimli öğrenme için kullanılan dağılımdan bağımsız bir öğrenme yöntemidir. İç karar düğümlerinden ve uç yapraklardan oluşur. Her düğüm bir denetim işlevi gerçekleştirir. Bu denetim işlevinin sonucunda göre dallardan biri seçilir. Yaprakta yazılı değer de çıktıyı oluşturur. Çıktı sınıflandırma problemi ise sınıf etiketi, bağlanım ise sayısal bir değer demektir. Karar ağacı öğrenmek veri kümesine bağlı olarak bir ağaç oluşturmak demektir. Ağaç kullanılan verinin yapısındaki karmaşıklığa göre büyür [24].

Karar ağaçları oluşturularak, karmaşık bir işlev, basit karar yapılarına dönüştürülür. Bir veri kümesinden Şekil 3.1'deki gibi birden fazla karar ağacı oluşturulabilir. Bu durumda boyutu küçük olan ağaç tercih edilir. Düğüm sayısı ve düğümlerdeki karar işlevlerinin karmaşıklığı boyutu belirler [24]. Tek değişkenli karar ağaçlarında her iç



Şekil 3.1: Karar ağacı oluşturma.

düğümde yalnız bir değişken kullanır. Değişken ayrık ise ve n farklı sonucu varsa, girdi uzayını n parçaya böler. Değişken sürekli ise, belirli bir eşik değerine göre uzayı iki parçaya böler. Bölme işlemine bir düğüme ulaşan örneklerin saflığına göre karar verilir. Saflık ölçütü olarak entropi (düzensizlik) kullanılabilir. p_m^i, m düğümüne ulaşan örnekler içinde C_i sınıfının olasılığı olmak üzere düzensizlik denklem (3.1)'deki gibi ölçülür [24].

$$I_m = -\sum_{i=1}^{K} p_m^i log_2 p_m^i \tag{3.1}$$

Bir m düğümüne düşen örnekler saf değilse bölme işlemi uygulanır. Bunun için girdilerden birini seçmek gerekir. Bütün olası girdiler için toplam saflık değeri bulunur. N_m, m düğümüne ulaşan örnekleri; N_{mj} m düğümünden j dalına düşen örnekleri; p_{mj}^i

de, m düğümünden j dalına düşen örnekler içinde C_i sınıfının olasılığını göstermek üzere, toplam saflık değeri denklem 3.2'deki gibi ölçülür. Toplam saflık değeri en düşük çıkan girdi bölme işleminde kullanılmak üzere seçilir. [24].

$$I'_m = -\sum_{j=1}^{n} \frac{N_{mj}}{N_m} \sum_{i=1}^{K} p^i_{mj} log_2 p^i_{mj} \quad \textbf{(3.2)}$$

Ağaç oluşturma işlemi sonrasında, karmaşıklığı azaltmak için, gereksiz dallar budanabilir [24].

3.2.2 Rastgele Orman

Rastegele orman içinde birçok sınıflandırma ağacı barındıran ve doğruluk değerini (accuracy) çok fazla arttıran bir yöntemdir [25]. Yeni bir örnek verildiğinde girdi vektörü ormandaki her bir ağaç tarafından ayrı ayrı sınflandırılır. Buna ağaç oylaması da denir. Sınıf etiketi oy çoğunluğu esasına göre belirlenir. Ağaç sayısı ayarlanabilir.

M uzaydaki örnek sayısı olmak üzere, her ağaç N kadar örnekten yerine koyma yöntemiyle (with replacement) rastgele örnekleme yapar. Yerine koyma yöntemi ile her örneklemede seçilme ihtimalleri aynı kalır. Seçilen örnekler o ağacın eğitim kümesini oluşturur. M girdi değişkenlerinin sayısını göstermek üzere, $m < M$ değişken (öznitelik) her ağaç için rastgele seçilir. m bütün orman için sabit kalır. Her ağaç budama işlemi olmadan mümkün olduğunca büyütülür.

Orman hata oranı herhangi iki ağaç arasındaki korelasyona ve her ağacın gücüne bağlıdır. İki ağaç arasındaki korelasyon arttıkça orman hata oranı da artar. Ormandaki her bir ağacın gücü arttıkça orman hata oranı düşer. Düşük hata oranına sahip ağaçlar güçlü sayılır [31].

Rastegele orman yöntemi büyük veri tabanlarında ve yüksek sayıda girdi değişkeni olan verilerde iyi performans gösterir. Sınıflandırmada değişkenlerin önemini kestirir. Bunun yanında, büyük bir kısmı eksik verilerle kesinliği yüksek tahminler geliştirir. Orman ilerde kullanılmak üzere muhafaza edilebilir.

Rastgele orman metodu önem (importance) işlevi sayesinde girdilerin önem derecelerini ölçer. İki türlü önem ölçümü yapar. İlki kesinlik değerindeki ortalama azalış, ikincisi düğüm saflık değerindeki ortalama azalış, başka bir ifade ile ortalama gini azalışıdır. İlki şöyle hesaplanır. Her ağaç için, verinin torba-dışı (out-of-bag) kısmı üzerinden tahmin hatası kaydedilir. Aynı işlem bütün kestirici değişkenlerin yeri

değiştirildikten sonra yapılır. İkisi arasındaki farkın ortalaması bütün ağaçlar üzerinden alınır ve normalleştirilir. İkincisi, belirli bir değişken üzerinde bölmeden kaynaklanan düğüm saflıklarındaki toplam düşüşün, bütün ağaçlar üzerinden ortalaması alınarak hesaplanır. Eğer istenmişse bütün girdi değişkenleri için bu değerler hesaplanır. Kesinlik azalış değeri bir özniteliğin, diğer özniteliklerle etkileşimini de düşünerek modeldeki önemini verir. Gini azalış değeri ise tek başına bir özniteliğin ayırmadaki gücünü ölçer.

Rastgele orman metodu örnekleme boyutu parametresi model geliştirilirken her sınıftan ne kadar örnekleme yapılacağını bildirir. Örnekleme asıl veriden yerine koyma şeklinde yapılır. Bu parametre rastgelelik etkisi sağlar. Bu parametre ile ormandaki her ağaç asıl verinin farklı bir yüzünü görür. Dengesiz sınıf dağılımına sahip veri kümelerinde model geliştirilirken dengesizlik problemini aşmaya ve modelin performansını arttımaya yardımcı olur. Örnekleme boyutu parametresi büyüdükçe rastgelelik azalır, veriyi ezberleme riski artar. Çok küçük seçildiği takdirde ormandaki ağaçların varyansı büyür. Veriyi ezberleme riskini azaltır ama genelde performans üzerinde kötü etki yapar.

3.2.3 Yapay Sinir Ağları

Sinir ağları işlemci birimi olan ve paralel çalışan çok sayıda nörondan ve aralarındaki bağlantıyı sağlayan çok sayıda sinapstan oluşur. Çok katmanlı algılayıcılar da sınıflandırmada ve bağlanımda kullanılabilen yapay sinir ağlarıdır [24].

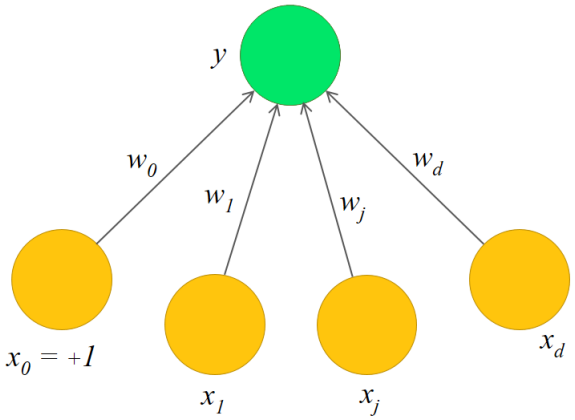
Algılayıcı, yapay sinir ağlarında temel işlemci birimidir (şekil 3.2). Girdileri ve çıktıları vardır. $x_j, j = 1,...,d$ girdi birimlerini gösterir. x_0 her zaman 1 değerini alan ek girdidir. w_j, x_j girdi biriminin ağırlığı, y de çıktı birimidir [24]. y çıktı birimi en basit durumda girdilerin ağırlıklı toplamları olarak hesaplanır (denklem 3.3) [24].

$$y = \sum_{j=1}^{d} w_j x_j + w_0 \tag{3.3}$$

Girdiler ve ağırlıklar vektör olarak yazıldığında çıktı iç çarpım olarak da tanımlanabilir (denklem 3.4) [24].

$$y = w^T x \tag{3.4}$$

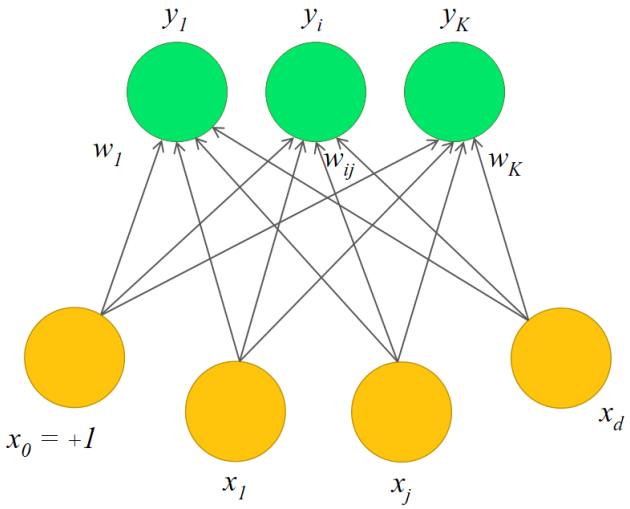
Bu durumda $x = [1, x_1, ..., x_d]^T$ ve $w = [w_0, w_1, ..., w_d]^T$ olur.



Şekil 3.2: Algılayıcı.

Tek girdili ve tek çıktılı algılayıcı, $y = wx + w_0$, doğru gerçekler ve çıktının işaretine göre doğrusal ayrılabilen iki sınıfı seçebilir [24].

$K > 2$ sınıfı seçmek için, şekil 3.3'deki gibi K çıktı gerekir. Bu da K algılayıcı manasına gelir [24]. Her algılayıcının kendi ağırlık vektörleri vardır ve çıktı girdilerin



Şekil 3.3: K paralel algılayıcı. $x_j, j = 0,...,d$ girdileri, $y_i, i = 0,...,K$ çıktıları, w_{ij} de x_j girdisinden y_i çıktısına olan bağlantının ağırlığını ifade eder.

ağırlıklı toplamı olarak ifade edilir (denklem 3.5). W, $K \times (d+1)$ boyutunda, satırları algılayıcıların bağlantı ağırlıkları, w_{ij}'lerden oluşan bir ağırlık matrisidir [24].

$$y_i = \sum_{j=1}^{d} w_{ij} x_j + w_{i0} = {w_i}^T \qquad \textbf{(3.5)}$$
$$y = Wx$$

Sınıflandırmada, çıktı değeri en yüksek olan sınıf seçilir (denklem 3.7). Yalnız sınıf etiketleri değil, sonsal (posterior) olasılıklar da gerekli ise, diğer sınıfların çıktı değerlerinin de muhafaza edilmesi gerekir. Bu ağırlıklı toplamlar ve eşiksiz en büyük işlev hesaplama şeklinde iki adımda gerçeklenen tek bir çıktı katmanı ile sağlanır (denklem 3.6) [24].

$$o_i = {w_i}^T x$$
$$y_i = \frac{\exp o_i}{\sum_k \exp o_k} \tag{3.6}$$

$$\text{seç } C_i \text{ eğer } y_i = \max_k y_k \tag{3.7}$$

Algılayıcının eğitilmesinde genellikle çevrimiçi öğrenme kullanılır. Çevrimiçi öğrenmede bütün öğrenme uzayı yerine örnekler tek tek işlenir. Her örnekte ağın parametreleri güncellenir. Hata fonksiyonu da tekil örnek için tanımlanır.

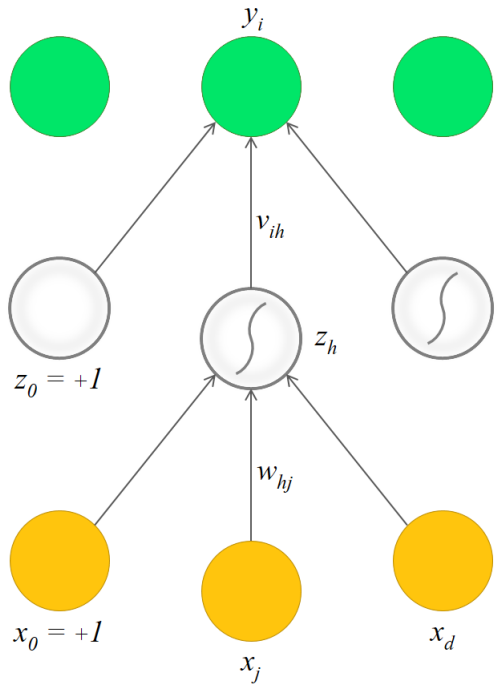
$K > 2$ sınıf için, çıktılar eşiksiz en büyük işlev ile hesaplanır (bkz. denklem 3.6). Çapraz düzensiziği ise denklem 3.8'deki gibi hesaplanır. (x^t, r^t) örneğinde *eğer* $x_t \in C_i$ *ise* ${r_i}^t = 1$ *değilse* ${r_i}^t = 0$ olur. ${r_i}^t$ istenen, ${y_i}^t$ ise gerçek çıktıdır.

$$E^t(\{w_i\}_i | x^t, r^t) = -\sum_i {r_i}^t \log {y_i}^t \tag{3.8}$$

Eğim iniş (gradient descent) kullanarak çevrimiçi güncelleme kuralı denklem 3.9'deki gibi yazılır. Böylece her örnekte ağ parametreleri güncellenir ve model gelişir [24].

$$\Delta w_{ij}^t = \eta (r_i^t - y_i^t) x_j^t \tag{3.9}$$
$$i = 1, ..., K$$
$$j = 0, ..., d$$

Ayırtacın doğrusal olmadığı durumlar ancak çok katmanlı algılayıcılar ile gerçeklenebilir (şekil 3.4) [24]. Çoklu katmanda x_j girdileri, tek katmanlıda olduğu gibi y_i çıktıları yerine, aradaki saklı birimleri beslerler. Saklı katmanda bulunan ve daima $+1$ değeri alan z_0 ek birimi ile birlikte x_j'den alınan girdiler genişleterek çıktı katmanı birimleri, y_i'e aktarılır. Bir saklı katman olduğunda çıktıyı hesaplarken önce x_j girdilerinin ağırlıklı toplamına *S işlemi* uygulanarak z_h saklı birimlerinin değerleri bulunur (denklem 3.10), daha sonra z_h girdilerinin ağırlıklı toplamı ile y_i çıktı değerleri



Şekil 3.4: Çok katmanlı algılayıcı. $x_j, j = 0,...,d$ girdileri; $z_h, h = 1,...,H$, saklı birimleri; $y_i, i = 0,...,K$ de çıktıları ifade eder. z_0 saklı katmandaki ek girdidir. w_{ij} ve v_{ij} sırasıyla birinci ve ikinci katmandaki ağırlık parametreleridir.

elde edilir (denklem 3.11) [24].

$$z_h = sigmoid(w_h^T x) = \frac{1}{1+\exp[-(\sum_{j=1}^{d} w_{hj} x_j + w_{h0})]} \tag{3.10}$$

$$y_i = v_i^T z = \sum_{h=1}^{H} v_{ih} z_h + v_{i0} \tag{3.11}$$

Çok katmanlı algılayıcıda çıktı girdinin doğrusal olmayan bir işlevi biçiminde olduğundan birinci katmandaki w_{ij} ağırlıkları için eğim hesaplanırken zincir kuralı kullanılır (denklem 3.12) [24].

$$\frac{\partial E}{\partial w_{hj}} = \frac{\partial E}{\partial y_i} \frac{\partial y_i}{\partial z_h} \frac{\partial z_h}{\partial w_{hj}} \tag{3.12}$$

$K > 2$ sınıf olduğu durumda, güncelleme kuralı denklem 3.13'deki hata işlevi üzerinden, eğim inişle (gradient descent) denklem 3.14'deki gibi türetilir [24].

$$E(W,V|X) = \sum_t \sum_i r_i^t \log y_i^t \tag{3.13}$$

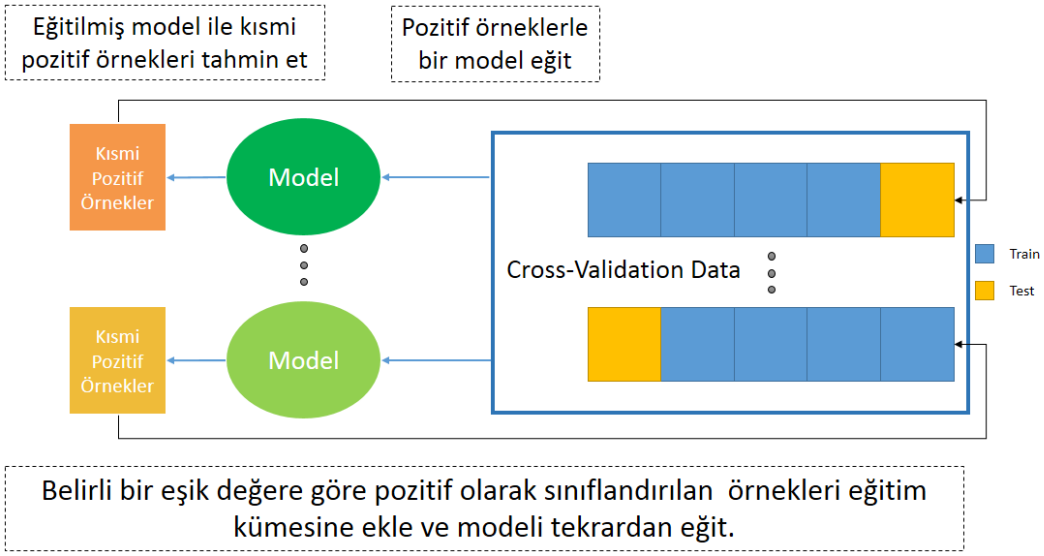
$$\Delta v_{ih} = \eta \sum_t (r_i^t - y_i^t) z_h^t$$

$$\Delta w_{hj} = \eta \sum_t \left[\sum_i (r_i^t - y_i^t) v_{ih} \right] z_h^t (1 - z_h^t) x_j^t \tag{3.14}$$

3.3 Beraber Öğrenme

Veri kümemiz biri HIV-1'e diğeri insana ait olan protein çiftlerinden oluşur. Yapay öğrenme yöntemleri ile tahmin edilmek istenen, HIV-1 ile insan proteinleri arasında hangilerinin etkileşime girdiğidir. Yani ikili sınıflandırma söz konusudur. Yapay öğrenmede kullanılan birçok veri kümesinden farklı olarak, kullanılan bu veri kümesinde "kısmi pozitif" kavramı yer alır. Bu çalışmanın temelini de, bu kısmi pozitiflerin nasıl daha etkili kullanılabileceği konusu oluşturur.

Kısmi pozitiflerin çözüme katılmasında birkaç farklı strateji düşünülebilir. Bunların ilk akla geleni, bütün kısmi pozitiflere pozitif gibi davranmak ve eğitim kümesine dahil etmektir. Kısmi pozitiflerlere tümden pozitif gibi davranmak, fazla gürültüye sebep olur ve kesinlik (precision) değerini çok fazla düşürür. Onun yerine bizim önerdiğimiz strateji şu şekildedir. Kısmi pozitif protein çiftleri dışarda tutularak, sadece altın standart pozitiflerle bir model eğitilir. Eğitilen bu modelden kısmi pozitifleri tahmin etmesi istenir. Tahmin edilen kısmi pozitiflerden belli bir eşik değerini geçenler eğitim kümesine pozitif olarak eklenir. Model tekrardan eğitilir ve kısmi pozitiflerden kalanını tahmin etmesi istenir. Bu işlem eğitim kümesine eklenecek kayda değer miktarda kısmi pozitif örnek kalmayıncaya kadar devam ettirilir (Şekil 3.5).



Şekil 3.5: Kısmi pozitifleri çözüme dahil etme.

Proje uygulanma aşamasında Weka [32] ve, R [33] araçları kullanıldı. R, hızlı çalışan esnek bir komut dilidir. Etkili bir veri yönetimi ve depolama imkanı sağlar.

Eklenti paketleri ile işlerliği arttırılabilir. Yoğun hesaplama gerektiren işlerde fortran ve C kodları ile bağlantı kurulabilir ve çalışma zamanında çağrılabilir. Weka yapay öğrenme yöntemlerini barındıran Java tabanlı bir araçtır. Arayüzden ve kod içerisinden çalıştırılma imkanını sunar. Veriyi ön işleme, sınıflandırma, regresyon, kümeleme ve görselleştirme gibi işlemler için hazır araçlar barındırır. Weka ve R genel kamu lisansına (GNU general public license) sahip yazılımlardır. Bu çalışmada testler çoğunlukla R üzerinden yapıldı.

4. DENEYLER

Bu tezde yapay öğrenme yöntemi olarak rastgele orman ve çok katmanlı algılayıcılar yöntemleri kullanıldı. Başarı oranını en yüksek düzeye çıkarmak için farklı yaklaşım ve girdi değerleri kullanılarak birçok deney yapıldı.

Farklı girdi değerleri ile rastgele orman (RO) ve yapay sinir ağları (YSA) yapay öğrenme yöntemleri denendi. Elde edilen sonuçlar kullanılan değerlendirme ölçütleri ışığında karşılaştırıldı. RO yöntemi, YSA yöntemine göre daha iyi performans sağladığı için, ileri düzey yaklaşımlarda tercih edildi.

Tezin bu bölümde kullanılan değerlendirme ölçütlerinden ve yapılan deneylerin sonuçlarından bahsedildi. Deneyler üç farklı başlık altında toplandı.

4.1 Değerlendirme Ölçütleri

Sınıflandırıcı performansının ölçülmesinde çeşitli ölçütler kullanılır. Bütün bu değerlendirme ölçütleri tabanda Tablo 4.1'de gösterilen doğru pozitif tanımlarına dayanır. $\hat{P}(C_1|x)$ pozitif sınıfın olasılığını ve $\hat{P}(C_2|x)$ negatif sınıfın olaslığını

Çizelge 4.1: Tahmin ve gerçek değer üzerinden doğru pozitif tanımı.

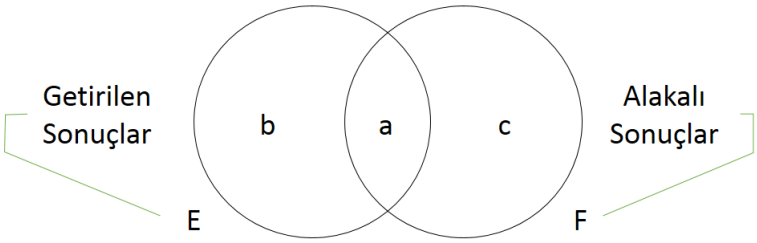
		Tahmin	
		+	**-**
Gerçek	**+**	Doğru Pozitif	Yanlış Negatif
	-	Yanlış Pozitif	Doğru Negatif

göstermek üzere $\hat{P}(C_2|x) = 1 - \hat{P}(C_1|x)$ eşitliği vardır. $\hat{P}(C_1|x) > \Theta$ olduğunda pozitif sınıf seçilsin. farklı Θ değerlerine göre farklı sonuçlar elde edilir.

Farklı Θ değerlerine göre farklı doğru-pozitif ve yanlış-pozitif oranları elde edilir. Bunlar birleştirilerek ROC (receiver operating characteristic) eğrisi elde edilir. ROC eğrisi altında kalan alana AUC (area under the curve) denir. Bir sınıflandırıcı, ROC eğrisi sol üste ne kadar yakınsa, başka bir ifade ile AUC değeri bire ne kadar yakınsa,

o kadar tercih edilir. İdeal olanı, doğru-pozitif oranı bir iken yanlış-pozitif oranının sıfır olduğu sınıflandırıcıdır.

Bir anahtar kelime ile bir veritabanına sorgu atıldığında dönen sonuçların bir kısmı aramamızla alakalı olabilir. Bunlar doğru-pozitif olanlardır. Ancak bütün alakalı sonuçlar gelmemiş olabilir. Bunlar yanlış-negatif olanlardır. Bazı sonuçlar ise alakasız olduğu halde getirilmiş olabilir. Bunlar da yanlış-pozitif sonuçlardır. Bu anlatım Şekil 4.1'deki gibi görsel olarak özetlenebilir. Şekil 4.1 üzerinden kesinlik (precision) ve



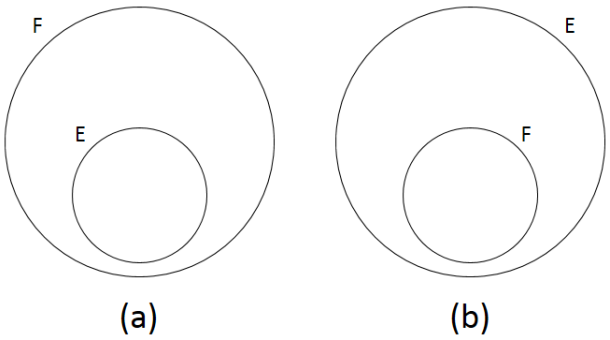
Şekil 4.1: E, getirilen; F, alakalı sonuçlar kümesini gösterir. Kümelerde bulunan alanlardan a, getirilen alakalı sonuçları; b, getirilen alakasız sonuçları; c, getirilmeyen alakalı sonuçları gösterir.

anma (recall) denklem 4.1 ve denklem 4.2'deki gibi hesaplanır.

$$\text{kesinlik} = \frac{a}{a+b} \tag{4.1}$$

$$\text{anma} = \frac{a}{a+c} \tag{4.2}$$

Şekil 4.2 (a)'da alakalı sonuçlar kümesi getirilen sonuçlar kümesini kapsar. Getirilen bütün sonuçların alakalı olduğu manasına gelir. Bu durumda kesinlik bir olur. Ancak getirilmeyen alakalı sonuçlar bulunabilir. Şekil 4.2 (b)'de ise tam tersi şekilde getirilen sonuçlar kümesi, alakalı sonuçlar kümesini kapsar. Bütün alakalı sonuçlar getirilmiş manasına gelir. Bu durumda da anma (geri-çağrışım) bir olur. Ancak getirilen sonuçlar alakasız olanlar da bulunabilir. Bu tezde kullanılan değerlendirme ölçütleri MAP (mean average precision), PRBE (precision-recall break-even point) ve AUC (area under the curve) ölçütleridir. MAP ayırt etme ve kararlılığı ölçme bakımında yakın zamanda kullanılan ölçütlerden biridir [26]. Kesinlik / anma eğrisini özetlemek için kullanılır. Kesinlik / anma eğrisinde, birbirine eşit uzaklıkta 10 farklı anma noktasına karşılık gelen kesinlik değerlerinin ortalaması alınarak hesaplanır. PRBE kesinlik ve anma birbirine eşit olduğu değerdir. Başka bir ifade ile kesinlik / anma grafiğinin



Şekil 4.2: (a)'da alakalı sonuçlar kümesi getirilen sonuçlar kümesini kapsar. Bu durumda kesinlik bir olur. (b)'de getirilen sonuçlar kümesi alakalı sonuçlar kümesini kapsar. Bu durumda da anma bir olur.

köşegeni kestiği yerdir. Bir ya da birden fazla kesme noktası olabilir. Bu durumda en son kesme değeri alınır. AUC, ROC eğrisinin altında kalan alana eşittir. ROC eğrisi, farklı ayrım eşik değerleri için yanlış pozitif oranlarına karşılık gelen doğru pozitif oranlarından oluşur. AUC, ROC eğrisini özetleyen ortalama bir performans değeri verir. Çizge olarak doğru-pozitif/yanlış-pozitif, kesinlik/anma, hassaslık/özgüllük, yoğunluk ve kutu çizgeleri kullanıldı. Ayrıca pozitif ve negatif sınıfların yoğunluk çizelgeleri de sonucu anlamaya ve yorumlamaya yardımcı olması amacıyla kullanıldı.

4.2 Deneyler ve Sonuçları

Deneylerde 5-kat çarpraz-sağlama (5-fold cross-validation) kullanıldı. Deneyler en az 10 kere çalıştırıldı ve elde edilen sonuçların ortalaması alındı. 5-kat ÇS verisi oluşturulurken şu şekilde oluşturuldu. Veri kümesinde bulunan pozitif ve negatif örnekler 5 parçaya ayrıldı. Daha sonra her kat bir parça pozitif ve bir parça negatif içermek üzere oluşturuldu. Böylece pozitif ve negatif örnekler katlar oluşturulurken eşit bir şekilde dağıtılmış oldu.

Pozitif ve kısmi pozitif örnekler dışında kalan 352328 örnekler olası negatif örnekleri oluşturur. Ancak 158 altın standart pozitif örneğin yanında bu sayı çok fazladır, ve yapay öğrenmede dengesizlik sorununa yol açar. Bunun önüne geçmenin bir yolu da örneklemedir. HIV-1 ve insan protein çiftleri arasındaki etkileşimin yaklaşık olarak 1/100 oranında olduğu [12] göz önünde bulundurularak olası negatif örneklerden 16000 rastgele örnekleme yapılarak negatif örnek kümesi oluşturuldu ($158 \times 100 =$

15800). Bu işlem her çalıştırmada rastgele olarak gerçeklenerek, şanstan oluşabilecek sonuçların önüne geçildi.

WEKA'da örnekleme işlemi için *SpreadSubSample* filtresi kullanıldı. Bu filtre kullanılarak, her bir sınıftan ne kadar örnekleme yapılacağı belirlenebilir. Dağılım yayılması (Distribution Spread) parametresi 0 olduğunda en yüksek miktarda yayılım gösterir. Yani her sınıf etiketinden bütün örnekleri seçer. Bu parametre 1 olduğunda dağılım tek düze (uniform) olur. Bu durumda her sınıftan eşit sayıda örnek seçilir ve en az sayıda örneği bulunan sınıf, sınıflardan seçilecek örnek sayını belirler. 10 olduğunda $10:1$ oranı sağlanır. Bu çalışmada kullanılan oran $1/100$ olduğu için parametre 100 olarak ayarlandı.

Rastgele orman (RO) ve yapay sinir ağları yöntemleri (YSA) hem WEKA hem de R ile denendi. Weka ile yapılan denemelerde Java kod ortamı ve Weka arayüzü kullanıldı. Hem daha esnek olması, hem de daha hızlı çalışmasından dolayı R dili tercih edildi. R, paketleri sayesinde farklı yapay öğrenme yöntemleri için uygun bir geliştirme ortamı sağlar. Rastgele orman yöntemi için orjinali Breiman [25] tarafından yazılan algoritmanın R uyarlaması olan *randomForest* R paketi kullanıldı [27]. YSA için ise *monmlp* R paketi kullanıldı.

Öncelikli olarak sade yaklaşım denendi. Sade yaklaşımda kısmi pozitif örnekler kullanılmadan, sadece altın standart pozitif ve rastgele seçilen negatif örnekler kullanılarak model geliştirildi. Bu yaklaşım ile bazı girdiler değiştirilerek elde edilen sonuçlar gösterildi.

İkinci olarak bütün kısmi pozitiflerin, pozitif sayılarak eğitim kümesine dahil edildiği yaklaşım denendi. Bu yaklaşımda kısmi pozitifler çapraz-sağlama verisinde yalnız eğitim kümesine dahil edildi. Çapraz-sağlama verisi oluşturulurken pozitif ve negatif örneklerden oluşan veri kümesi n parçaya ayrılır. n farklı kat oluşturulur. Her kata n parçanın $n-1$ tanesi eğitim, 1 tanesi test olarak eklenir. Bütün kısmi pozitiflerin eklendiği yaklaşımda, ayrıca kısmi pozitifler de n parçaya ayrıldı ve katlardaki eğitim veri kümelerine dağıtıldı. Böylece pozitif olarak eklenen kısmi pozitifler sadece model eğitilirken kullanıldı, testler sonuçları ise bilinen pozitif örnekler üzerinden elde edildi. Bu yaklaşım için de girdiler değiştirilerek elde edilen farklı sonuçlar gösterildi.

Son olarak kısmi pozitiflerin eğitim kümesine dahil etmeden önce, bir ön işlemden geçirildiği yaklaşım denendi. Bu yaklaşım ile altın standart pozitif ve örneklenen negatif protein çiftleri ile bir model geliştirildi. Yöntem bölümünde bahsedildiği şekilde önce kısmi pozitifler eğitilen bu model ile tahmin edildi. Belirli bir eşik değerini geçen örnekler eğitim kümesinde dahil edildi. Önceki yaklaşımda açıklandığı şekilde, eklenen kısmi pozitifler çapraz-sağlama verisinde yalnız eğitim kümesine dahil edildi. Genişletilmiş eğitim kümesi ile yeni bir model geliştirildi ve kalan kısmi pozitif örnekleri tahmin etmesi istendi. Aynı şekilde eşik değeri geçen kısmi pozitif örnekler pozitif olarak eğitim kümesine eklendi. Bu işlem pozitif olarak ayrılan kayda değer kısmi pozitif kalmayana dek devam ettirildi.

Deney açıklamalarında kullanılan varsayılan ayarlar ifadesi deneyde kullanılan bazı temel girdilerin aldığı varsayılan değerleri ifade eder. Varsayılan ayarlar şunları kapsar. Rastgele orman yapay öğrenme yönteminde ağaç sayısı 500, örnekleme boyutu bütün sınıflar için en küçük sınıf boyutu olarak ayarlandı. Örnekleme boyutu için bu ayar, k en küçük sınıf boyutunu göstermek üzere, kısaca (k,k) olarak da ifade edildi. Başlangıçta oluşturulan veri kümesi, negatif sınıftan alınan 16000 rastgele örnekten ve pozitif sınıftan alınan mevcut 158 örnekten oluşur. Çapraz-sağlama verisi 5-kat olarak oluşturuldu. Varsayılan ayarlar ifadesine ek olarak belirtilen ifadelerle, o deneye özgü yapılan değişiklikler veya eklemeler anlatıldı. Deneyler 10 kere tekrarlandı. Böylece şanstan doğabilecek sonuçların önüne geçildi. Aynı metod ve aynı ayarlar ile bir deney en az 10 kere tekrar edildikten sonra çıkan sonucun ortalamaları alındı.

RO yönteminde bulunan örnekleme boyutu girdisi, ağaçlar oluşturulurken sınıflardan yapılacak örneklemenin sayısını ayarlamaya yarar. Bu ayar ile kullanıldığında ağaçlar büyütülürken pozitif ve negatif sınıflardan en küçük sınıfın boyutuna göre örneklem alınır. En küçük sınıf boyutunun k olduğu varsayılırsa, ikili sınıftan oluşan problemimizde örnekleme boyutu parametresi [k, k] olarak verilir. Böylecek pozitif ve negatif sınıflardan ağaç oluşturulurken k örnek alınır. Yani boyutu küçük olan pozitif sınıfın bütün elemanları kullanılırken, negatif sınıftan k boyutta örneklem alınır. Bu girdi, dengesiz sınıf dağılımına dahip veri kümelerinde model geliştirilirken dengesizlik problemini aşmaya ve modelin performansını arttırmaya yardımcı olur.

4.2.1 Sade yaklaşım

Sade yaklaşımda kısmi pozitif örnekler kullanılmadı. Model geliştirilirken, altın standart pozitif örnekler ile belli sayıda örneklenen negatif örnekler kullanıldı.

Sade yaklaşımda sınıflandırma yöntemi olarak rastgele orman (RO) ve yapay sinir ağları (YSA) kullanıldı. RO gürültülü ve artıklı verilerde iyi performans sağlar [13]. Yapılan testlerin sonucuna göre de, RO protein-protein etkileşimini belirlemede YSA'ya göre daha iyi sonuç verdi (Çizelge 4.2). Bu sonuçlar 10 ayrı çalışmanın sonuçlarının ortalaması alınarak elde edildi. RO'da varsayılan ayarlar kullanıldı. YSA'da 1 gizli katman, 3 gizli birim kullanıldı. Küme sayısı parametresi (ensemble) olarak 20 verildi. Sade yaklaşımda RO ile yapılan 10 deneyin ayrıntılı sonuçları

Çizelge 4.2: Sade yaklaşımda, yapay sinir ağları ve rastgele orman yöntemleri kullanılarak yapılan deneylerin deneylerin AUC, PRBE ve MAP ölçütlerine göre ortalama sonuçları.

Ölçüt	Yapay Sinir Ağları	Rastgele Orman
AUC	0.890	0.922
PRBE	0.209	0.245
MAP	0.101	0.157

4.3'deki gibidir. Bu deneyde varsayılan ayarlar kullanıldı. Varsayılan ayarların neler olduğu daha önce belirtilmişti. İleride verilecek deney sonuçları bu sonuçlar ile kıyaslanabilir. Çizelge 4.3'deki sonuçların ortalama AUC, PRBE ve MAP değerleri

Çizelge 4.3: Sade yaklaşım ve varsayılan ayarlarda, rastgele orman yöntemi kullanılarak elde edilen 10 çalıştırmanın ayrıntılı sonuçları.

	Testler									
Ölçütler	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
AUC	0.919	0.920	0.925	0.920	0.919	0.925	0.922	0.922	0.923	0.922
PRBE	0.229	0.240	0.238	0.245	0.227	0.224	0.229	0.236	0.252	0.239
MAP	0.152	0.153	0.158	0.159	0.145	0.161	0.144	0.159	0.156	0.149

sırası ile 0.922, 0.236 ve 0.154 olur (4.4). Bu deneyde ağaç sayısı 500 ve örnekleme boyutu (sample size) en küçük sınıf boyutu olacak şekilde ayarlandı. Örnekleme boyutu girdisi, rastgele orman yönteminde ağaç oluşturulurken hangi sınıftan ne kadar örnekleme yapılacağını belirlemeye yarar. Problem iki sınıftan oluştuğu için bu girdi, ilki negatif, ikincisi pozitif sınıfı belirtmek üzere iki boyutlu bir sayı dizisi

Çizelge 4.4: Çizelge 4.3'deki sonuçların ortalama AUC, PRBE ve MAP değerleri. (m) ortalama, (s) standart sapmayı belirtir.

AUC(m)	AUC(s)	PRBE(m)	PRBE(s)	MAP(m)	MAP(s)
0.922	0.002	0.236	0.009	0.154	0.006

şeklindedir. Kullanılan veri kümesinde, pozitif sınıf boyutu negatif sınıfa göre çok küçük olduğundan pozitif sınıfın tamamı kullanıldı. Negatif sınıftan ise pozitif sınıfa oranla örneklem alındı. Örneğin, pozitif sınıfın boyutu *k*, çarpan *2* olduğunda örnekleme boyutu girdisi [*2k*, *k*] olur. Böylece pozitif sınıfın tamamı (*k*) alınırken, negatif sınıftan *2k* miktarında örneklem alınır. Bu çarpan 1, 2, 3, 4 ve 5 olacak şekilde test edildi (Çizelge 4.5). Mevcut ölçütler baz alındığında, bu sonuçlara göre

Çizelge 4.5: Sade yaklaşım ve varsayılan ayarlarda, örnekleme boyutu çarpanı 1'den 5'e kadar değiştirilerek yapılan deneylerin ortalama sonuçları. En son satırdaki sonuçlar bütün sınıflardan var olan bütün örnekler kullanılarak elde edildi.

Örnekleme	AUC		PRBE		MAP	
Boyutu Çarpanı	Ort.	Sd.	Ort.	Sd.	Ort.	Sd.
1	0.922	0.002	0.236	0.009	0.154	0.006
2	0.923	0.002	0.265	0.014	0.163	0.007
3	0.923	0.003	0.287	0.017	0.176	0.013
4	0.922	0.004	0.294	0.015	0.176	0.010
5	0.923	0.004	0.282	0.017	0.174	0.013
-	0.919	0.008	0.297	0.020	0.183	0.012

en iyi performans örnekleme boyutu çarpanı 4 alındığında elde edilir, yani ormandaki her bir ağaç oluşturulurken negatif sınıftan pozitif sınıf boyutunun 4 katı kadar örnekleme alındığında. Bu deneylerde örnekle boyutu parametresi değiştirilirken, diğer parametreler varsayılan ayarlarda kullanıldı. Varsayılan ayarlarda ağaç sayısı 500 idi. Son satırdaki sonuçlar orman oluşturulurken pozitif ve negatif sınıfın bütün örnekleri kullanılarak elde edildi. Bu durumda AUC değerinde düşüş olurken PRBE ve MAP değerlerinde yükselme oldu.

Ormandaki ağaç sayısının artmasıyla geliştirilen modelin, sapması (bias) yükselmeden, değişkesi (variance) azalma eğilimi gösterir [34]. Ormandaki her ağaç veriden rastgele örnekleme yaptığı için, ağaç sayısının arttırılması şanstan kaynaklı sonuçların azalmasında fayda sağlayabilir. Ağaç sayısı belli bir seviyeden sonra kayda değer bir

performans artışı sağlamaz. Ağaç sayısı 10, 25, 50, 100, 200, 300, 400, 500, 1000 ve 2000 olacak şekilde deneyler yapıldı. Her deney 10'ar defa çalıştırılarak ortalamaları alındı (Çizelge 4.6). Bu sonuçlara göre ağaç sayısı arttıkça performans artmaktadır.

Çizelge 4.6: Sade yaklaşım ve varsayılan ayarlarda, ağaç sayısı değiştirilerek elde edilen test sonuçlarının ortalama değerleri.

	AUC		PRBE		MAP	
Ağaç Sayısı	Ort.	Sd.	Ort.	Sd.	Ort.	Sd.
10	0.900	0.006	0.202	0.027	0.104	0.009
25	0.916	0.004	0.234	0.024	0.133	0.012
50	0.918	0.003	0.245	0.033	0.140	0.011
100	0.919	0.004	0.243	0.030	0.143	0.008
200	0.921	0.002	0.249	0.029	0.149	0.012
300	0.922	0.002	0.250	0.019	0.155	0.010
400	0.921	0.003	0.249	0.017	0.157	0.010
500	0.920	0.002	0.236	0.022	0.149	0.009
1000	0.922	0.002	0.255	0.028	0.164	0.014
2000	0.920	0.003	0.252	0.012	0.160	0.008

En iyi performans, ağaç sayısı 1000 olduğunda elde edildi. Ağaç sayısı 2000 verilerek elde edilen test sonuçlarının kesinliği 1000'e göre daha fazladır. Ancak ağaç sayısı 2000 verildiğinde test süresi çok fazla artmaktadır.

Veri kümemizde 158 pozitif protein çifti bulunmaktadır. Bu rakam geriye kalan olası negatif örneklere göre çok azdır. Bundan dolayı olası negatif örneklerden testin en başında rastgele örnekleme yapıldı. Pozitif örneklerin sayısı kadar kapsayıcılığı yani temsil kabiliyeti de önemlidir. Bunu ölçmek için mevcut pozitif örneklerin sayısı düşürülerek deneyler yapıldı. Çizelge 4.7'daki sonuçlar pozitif örneklerin sayısı düşürülerek elde edildi. Bu işlem çapraz-sağlama verisi oluşturulduktan sonra yapıldı. Her kat için yalnız eğitim kümesinde bulunan pozitif örnekler belli oranlarda ve rastgele seçilerek azaltıldı, test kümesine ise dokunulmadı. Varsayılan ayarlarda, RO'daki örnekleme boyutu parametresi [k, k] olarak belirlenmişti. Bu deneyde de bu ayarlar kullanıldı. Örnekleme boyutu bu şekilde verildiğinde en küçük sınıf boyutu bütün sınıflar için örnekleme sayısını belirler. Bundan dolayı eğitim kümesindeki pozitif örneklerin sayısının düşürülmesi ağaç oluşturulurken negatif sınıftan yapılacak örnekleme sayısını da düşürür.

Çizelge 4.7: Sade yaklaşım ve varsayılan ayarlarda, pozitif örneklerin sayısının belli oranlarda azaltılması ile elde edilen test sonuçlarının ortalama değerleri. Yüzde (%), çapraz-sağlama verisindeki her katın eğitim kümesinde bırakılan pozitif örneklerin yüzdesini, pozitif örnek sayısı ise sayısını gösterir. Pozitif örneklerin %25'i çıkarıldığında eğitim kümesinde %75 yani yaklaşık 94 pozitif örnek kalır.

Yüzde (%)	Pozitif Örnek Sayısı ($\sim$)	AUC		PRBE		MAP	
		Ort.	Sd.	Ort.	Sd.	Ort.	Sd.
75	94	0.919	0.003	0.243	0.025	0.145	0.008
50	64	0.915	0.002	0.221	0.018	0.138	0.006
33	42	0.916	0.003	0.230	0.024	0.135	0.012
25	32	0.912	0.004	0.222	0.023	0.132	0.010
20	25	0.909	0.004	0.209	0.026	0.119	0.012
10	13	0.905	0.004	0.202	0.024	0.111	0.015

Örnekleme boyutu parametresi $[500, k]$ olacak şekilde ayarlandığında, yani negatif sınıftan yapılacak örneklemenin boyutu 500'de sabitlendiğinde Çizelge 4.8'deki ortalama sonuçlar elde edildi. Örnekleme boyutu negatif sınıf için sabit verildiğinde

Çizelge 4.8: Sade yaklaşım ve varsayılan ayarlarda, pozitif örneklerin sayısının belli oranlarda azaltılması ile elde edilen test sonuçlarının ortalama değerleri. Örnekle boyutu parametresi [500, k] olarak ayarlandı. Negatif sınıftan alınacak örnekleme boyutu sabitlendi.

Yüzde (%)	Pozitif Örnek Sayısı ($\sim$)	AUC		PRBE		MAP	
		Ort.	Sd.	Ort.	Sd.	Ort.	Sd.
75	94	0.921	0.003	0.280	0.019	0.174	0.017
50	64	0.912	0.004	0.257	0.021	0.155	0.014
33	42	0.908	0.006	0.262	0.020	0.154	0.014
25	32	0.899	0.004	0.246	0.022	0.144	0.009
20	25	0.896	0.006	0.246	0.026	0.136	0.014
10	13	0.880	0.005	0.225	0.024	0.123	0.017

de en küçük sınıf boyutuna göre ayarlandığında da, pozitif örneklerin sayısı çeyreğine kadar düşürüldüğü durumda bile ortalama ölçüt değerlerinde temel çalıştırma sonuçlarına göre çok fazla fark bir olmadığı görüldü. Bu durumun sebebi pozitif örneklerin kapsayıcılığının düşük olmasıdır. Kullandığımız veri kümesinde pozitif örneklerin az bir kısmı bütün pozitiflerin karakterini sergilemeye yetmektedir.

Rastgele orman yöntemi girdi değişkenlerinin önemini hesaplamaya yarayan önem işlevine sahiptir. Bu işlev ile her bir değişkenin, sınıflara göre ham önem değerleri,

kesinlik ortalama azalış değerleri ve gini ortalama azalış değerli hesaplanır. Çizelge 4.9'de sade yaklaşım ve varsayılan ayarlarla geliştirilen modellerdeki ortalama önem değerleri gösterildi. Kesinlik ortalama azalışı girdi değişkenlerinin modeldeki önemini

Çizelge 4.9: Sade yaklaşım ve varsayılan ayarlarla yapılan 10 deneyin sonucuna göre girdi değişkenlerinin ortalama önem değerleri. Öznitelik açıklamaları için bkz. Bölüm 3.1.

Öznitelikler	Doğruluk Ortalama Azalışı	Öznitelikler	Gini Ortalama Azalışı
V15	1.17e-02	V9	18.77
V3	2.92e-03	V16	17.65
V4	2.61e-03	V18	17.47
V2	1.94e-03	V15	11.96
V17	1.91e-03	V17	9.63
V10	1.50e-03	V6	6.87
V11	1.24e-03	V10	5.77
V1	6.11e-04	V3	5.44
V7	3.74e-04	V11	5.28
V5	2.54e-04	V4	5.02
V13	9.94e-05	V2	4.96
V14	6.15e-05	V7	4.73
V6	-5.02e-05	V8	4.52
V16	-7.42e-04	V5	2.19
V8	-8.96e-04	V14	2.01
V18	-1.40e-03	V12	1.94
V12	-3.48e-03	V13	1.39
V9	-1.00e-02	V1	0.58

gösterir. Diğer girdi değişkenleri ile etkileşimini Gini ortalama azalışı ise her bir değişkenin tek başına bölme gücünü ifade eder. Çizelge 4.9'deki değerler azalan şekilde sıralanmıştır. Gini ortalama azalış değerlerinde 9, 16, 18, 15 ve 17. girdi değişkenleri diğerlerine kıyasla yüksek değerlere sabittir. Kesinlik ortalama azalış değerlerinde ise 15. girdi değişkeni diğerlerine oranla çok yüksek bir değere sahiptir (V15/V3 = 4.01).

4.2.2 Bütün kısmi pozitiflerin pozitif sayıldığı yaklaşım

Bu yaklaşımda kısmi pozitif örneklerin tamamının eğitim kümesine pozitif olarak dahil edildiği durum denendi. Kısmi pozitifler çapraz-sağlama verisindeki katların eğitim kümelerine eşit olarak dağıtıldı, ancak test kümesine dahil edilmedi. Test kümesinde bulunan pozitif örnekler yalnızca doğruluğu deneysel olarak kanıtlanmış altın standart

pozitif örneklerden oluştu. Bu yaklaşıma kısaca "kısmi pozitif (KP) dahil" yaklaşımı denilsin.

Bütün kısmi pozitiflerin eğitim kümesine dahil edildiği deneyin varsayılan ayarlarda 10 kere çalıştırılması ile elde edilen ayrıntılı sonuçlar Çizelge 4.10'de gösterildi. Bu deney varsayılan ayarlarda 10 kere çalıştırıldı ve Çizelge 4.11'deki ortalama sonuçlar elde edildi. Bu deneyde kısmi pozitiflerin tümünün dahil edilmesi ile elde edilen

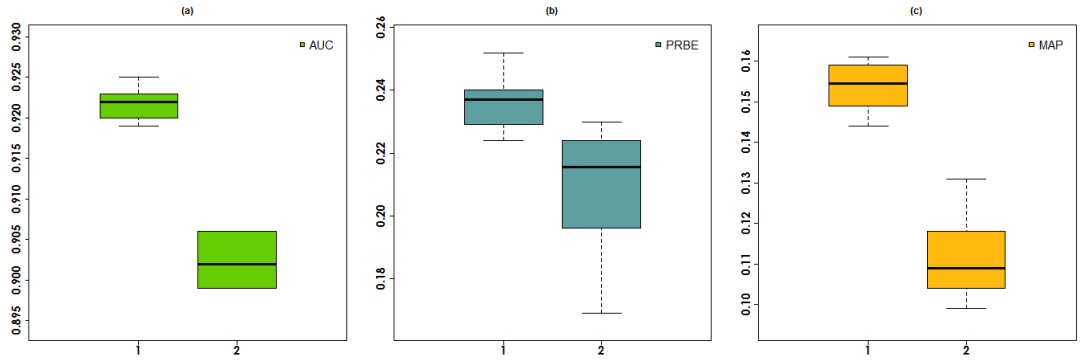
Çizelge 4.10: Kısmi pozitif örneklerin tamamımın pozitif sayılarak eğitim kümesine dahil edilmesi ile elde edilen ayrıntılı sonuçlar.

	Testler									
Ölçütler	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
AUC	0.906	0.903	0.899	0.906	0.899	0.899	0.902	0.902	0.901	0.906
PRBE	0.230	0.196	0.219	0.185	0.218	0.210	0.169	0.224	0.213	0.229
MAP	0.110	0.108	0.108	0.100	0.118	0.120	0.099	0.131	0.104	0.116

Çizelge 4.11: Çizelge 4.10'de gösterilen sonuçların ortalama değerleri.

AUC(m)	AUC(s)	PRBE(m)	PRBE(s)	MAP(m)	MAP(s)
0.902	0.003	0.209	0.020	0.111	0.010

PRBE ve MAP ölçüt değerleri sade yaklaşımla yapılan deneye göre düşmüştür. AUC ölçüt değeri ise pek değişmemiştir. İki deneyin ölçüt değerlerinin karşılaştırması Şekil 4.3'deki gibidir. Buna göre kısmi pozitiflerin tamamının eğitim kümesine eklenmesi, kullanılan ölçütlerin hiçbirinde modelin performansını arttırmadı. Bunun nedeni kısmi pozitif örneklerin deneysel olarak pozitifliği kanıtlanmamış ve dolayısı ile yüksek miktarda gürültü içerme ihtimalinin yüksek olmasıdır. Kısmi pozitiflerin iki gruptan oluştuğundan bahsedilmişti. Birinci grup, etkileşimi göstermesi bakımından ikinci gruba göre daha güçlü kelimelerle literatürde geçen protein çiftlerinden oluşur. Bütün kısmi pozitifler yerine sadece grup 1 kısmi pozitifler eğitim kümesine dahil edildiğinde Çizelge 4.12'deki ayrıntılı sonuçlar alındı. Benzer şekilde sadece grup 2 kısmi pozitif örnekleri eğitim kümesine dahil edildiğinde Çizelge 4.13'deki ayrıntılı sonuçlar alındı. Bütün kısmi pozitiflerin, sadece grup-1 ve sadece grup-2 kısmi pozitiflerinin eğitim kümesine pozitif olarak eklenmesi ile yapılan üç farklı deney kümesinin ortalama sonuçları Çizelge 4.14'de gösterildi. Bu sonuçlara göre AUC ölçütünde Grup-1 KP örneklerin eklenmesi diğerlerine göre daha iyi sonuç verirken, ancak PRBE ve MAP



Şekil 4.3: Sade yaklaşımla (1) ve KP örneklerin tamamının eğitim kümesine dahil edilmesiyle (2) yapılan testlerin AUC (a), PRBE (b) ve MAP (c) ölçüt değerlerine göre kutu çizim kullanılarak karşılaştırılması.

Çizelge 4.12: Kısmi pozitif sınıftan yalnız grup 1 örneklerin eğitim kümesine dahil edilmesi ile yapılan deneylerin ayrıntılı sonuçları.

	Testler									
Ölçütler	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
AUC	0.916	0.916	0.912	0.915	0.914	0.914	0.918	0.915	0.919	0.919
PRBE	0.181	0.184	0.166	0.195	0.176	0.146	0.183	0.165	0.188	0.186
MAP	0.117	0.106	0.103	0.106	0.109	0.099	0.110	0.110	0.111	0.109

ölçütlerinde daha kötü sonuç vermiştir. Grup-2 KP örneklerin eklenmesi ile tümünün eklenmesi yakın sonuçlar vermiştir.

Kısmi pozitif örnekler yerine negatif örnekler kullanıldığında Çizelge 4.15'deki ayrıntılı sonuçlar elde edildi. Negatif sınıftan KP sayısında örnekleme alındı ve KP örnekler gibi eğitim verisine pozitif olarak dahil edildi. Bu deneyle veri kümesindeki KP örneklerin olası negatif örneklerden farkı araştırıldı. Bu deneyde elde edilen ortalama sonuçlar Çizelge 4.16'de gösterildi. Bu sonuçlara göre kısmi pozitif örneklerin negatif örneklere göre pozitife daha yakın oldukları doğrulandı.

4.2.3 Kısmi pozitiflerin adım adım dahil edilmesi yaklaşımı

İlk olarak sade yaklaşımda model, kısmi pozitif örnekler hesaba katılmadan, sadece altın-standart pozitif örnekler ve olası negatif sınıftan belirli sayıda seçilen örnekler kullanılarak geliştirildi. İkinci olarak kısmi pozitiflerin tamamı eğitim kümesine dahil edilerek model geliştirildi. Bu bölümde adım-adım yaklaşımı denendi. Bu yaklaşımda, kısmi pozitif örnekler eğitim kümesine eklenmeden önce ön işlemden

Çizelge 4.13: Kısmi pozitif sınıftan yalnız grup 2 örneklerin eğitim kümesine dahil edilmesi ile yapılan deneylerin ayrıntılı sonuçları.

	Testler									
Ölçütler	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
AUC	0.917	0.915	0.916	0.920	0.918	0.918	0.915	0.914	0.916	0.919
PRBE	0.068	0.058	0.068	0.046	0.048	0.054	0.038	0.045	0.053	0.034
MAP	0.032	0.027	0.030	0.031	0.028	0.030	0.029	0.029	0.030	0.027

Çizelge 4.14: Kısmi pozitif örneklerin tamamının, sadece Grup-1 ve sadece Grup-2'den olanlarının eğitim kümesine eklenmesi ile yapılan deneylerin kümelerinin ortalama sonuçları.

	AUC		PRBE		MAP	
Yöntem	Ort.	Sd.	Ort.	Sd.	Ort.	Sd.
Bütün KP	0.902	0.003	0.209	0.020	0.111	0.010
Grup-1 KP	0.916	0.002	0.177	0.014	0.108	0.005
Grup-2 KP	0.901	0.002	0.211	0.023	0.118	0.013

geçirildi. Öncelikli olarak kısmi pozitifler kullanılmadan, sade yaklaşımda, bir model oluşturuldu. Oluşturulan bu model ile kısmi pozitif örnekler tahmin edildi. Tahmin oranına bakılarak belirli bir eşik değerin üstündeki kısmi pozitif örnekler eğitim kümesine dahil edilerek, yeniden model geliştirildi. Geliştirilen yeni model ile kalan kısmi pozitif örnekler tahmin edildi ve tekrardan eşik değeri geçenler eğitim kümesine dahil edildi. Bu işlem eğitim kümesine eklenecek kayda değer miktarda kısmi pozitif örnek kalmayana kadar tekrarlamalı olarak devam ettirildi. Bu yaklaşım ile, kısmi pozitiflerin en faydalı olacak şekilde modele dahil edilmesi amaçlandı. Kısmi pozitiflerin yüksek miktarda gürültü içermesinden dolayı, performanstaki olumsuz etkisinin önüne geçilmek ve modeli daha iyiye taşımak amaçlandı.

Adım-adım yaklaşımında, varsayılan ayarlar kullanılarak elde edilen ayrıntılı sonuçlar Çizelge 4.17'de gösterildi. Varsayılan ayarlarda eşik değer 0.9 olarak belirlendi. Eşik değer, tahmin edilen kısmi pozitif örneklerden eğitim kümesine eklenecek örnekleri seçmek için kullanıldı. Model tarafından verilen tahmin değeri, belirlenen eşik değerini geçen örnekler eğitim kümesine eklenmek üzere seçildi.

Bu deneyde, her adımda tahmin edilen ve eşik değeri geçen yeni kısmi pozitif örnekler eğitim kümesine eklenir ve model yeniden eğitilir. Çizelge 4.17'de gösterilen her adımda, yeni kısmi pozitiflerin eklenmesi ile genişletilmiş eğitim kümesi üzerinden

Çizelge 4.15: KP örnekler yerine negatif sınıftan aynı sayıda örneklenen örnekler kullanıldığında elde edilen ayrıntılı sonuçlar. Negatif örnekler pozitif gibi sayılarak ÇS eğitim kümelerine uygun biçimde dahil edildi.

	Testler									
Ölçütler	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
AUC	0.762	0.763	0.777	0.764	0.738	0.826	0.790	0.749	0.738	0.799
PRBE	0.089	0.108	0.077	0.083	0.000	0.129	0.051	0.098	0.058	0.101
MAP	0.038	0.043	0.043	0.034	0.024	0.053	0.035	0.038	0.024	0.051

Çizelge 4.16: KP örnekler yerine negatif sınıftan aynı sayıda örneklenen örnekler kullanıldığında elde edilen ortalama sonuçlar.

AUC(m)	AUC(s)	PRBE(m)	PRBE(s)	MAP(m)	MAP(s)
0.771	0.028	0.079	0.036	0.038	0.010

geliştirilen modelin test sonuçları yer alır. Numarasız olan ilk satırlar, kısmi pozitif örnekler kullanılmadan geliştirilen modelin test sonuçlarıdır. Bu deneye ait ölçüt değerlerinin ortalama sonuçları Çizelge 4.18'deki gibidir. Bu sonuçlara göre sade yaklaşımla geliştirilen model, kısmi pozitif örneklerin eklendiği durumlara göre daha iyi performans verir. Kısmi pozitifler örneklerin seçme işleminden geçirilerek, adım adım eklendiği durumda dahi sade yaklaşıma göre performans artışı sağlanamadı. Bu da gösteriyor ki kullandığımız veri kümesi için kısmi pozitiflerin herhangi bir şekilde eğitim kümesine dahil edilmesi performans üzerinde olumsuz etki yapar. Bunun bir nedeni kısmi pozitiflerin fazla gürültülü olmasıdır. Diğer bir nedeni de, daha önce belirtildiği gibi altın-standart pozitif örneklerin temsil niteliğinin düşük olmasıdır. Kısmi pozitifler eğitim kümesine eklenirken, test kümesi yalnız altın-standart pozitif örnekler içerir. Elde edilen bütün sonuçlar, uzmanlar tarafından onaylanmış pozitif örnekler kullanılarak elde edildi. Mevcut pozitif örneklerin kapsayıcılığının düşük olması durumunda, gerçekte pozitif olan bir örneğin model tarafından tanınmaması durumu çıkabilir. Bir başka neden de negatif örneklerin gerçekte negatif olduklarının kesin olmamasıdır. İki proteinin etkileşmediğini göstermek çok zordur. Bu nedenle negatif örnekler, veri kümesinde etkileştiği bilinen protein çiftleri dışında kalanlardan belirli miktarda örnekleme alınarak elde edilir. Dolayısı ile negatif örneklerin içinde bilinmeyen pozitif örnekler olabilir.

Çizelge 4.17: KP örneklerin eğitim kümesine adım-adım eklenmesi ile geliştirilen modelin ayrıntılı test sonuçları. Numarasız olan ilk satırlar, kısmi pozitifler eklenmeden önce oluşturulan modelin test sonuçlarını gösterir.

		Testler									
Ölçütler	**Adımlar**	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
AUC	-	0.924	0.918	0.922	0.924	0.920	0.922	0.921	0.920	0.922	0.924
	1	0.923	0.919	0.924	0.923	0.919	0.919	0.922	0.920	0.923	0.924
	2	0.924	0.918	0.923	0.923	0.918	0.920	0.922	0.919	0.923	0.923
	3	0.925	0.919	0.922	0.923	0.920	0.920	0.921	0.919	0.925	0.923
	4	0.924	0.918	0.922	0.924	0.920	0.921	0.921	0.918	0.922	0.923
	5	0.923	0.921	0.922	0.922	0.921	0.920	0.921	0.919	0.923	0.922
PRBE	-	0.229	0.245	0.239	0.221	0.273	0.247	0.237	0.242	0.260	0.237
	1	0.237	0.252	0.242	0.257	0.273	0.227	0.218	0.269	0.221	0.227
	2	0.247	0.239	0.231	0.244	0.265	0.214	0.218	0.230	0.248	0.217
	3	0.237	0.231	0.232	0.226	0.258	0.204	0.236	0.247	0.229	0.231
	4	0.239	0.252	0.236	0.213	0.250	0.232	0.217	0.238	0.216	0.217
	5	0.227	0.229	0.232	0.214	0.256	0.227	0.223	0.236	0.225	0.217
MAP	-	0.165	0.149	0.149	0.148	0.171	0.147	0.161	0.148	0.147	0.153
	1	0.159	0.160	0.151	0.155	0.166	0.150	0.147	0.157	0.150	0.150
	2	0.155	0.149	0.146	0.151	0.160	0.144	0.142	0.150	0.137	0.144
	3	0.158	0.145	0.147	0.142	0.158	0.134	0.139	0.147	0.137	0.138
	4	0.160	0.145	0.156	0.140	0.154	0.134	0.137	0.142	0.138	0.135
	5	0.152	0.150	0.144	0.143	0.149	0.140	0.135	0.146	0.135	0.134

Adım yaklaşımında çeşitli stratejiler ile girdi değerleri değiştirilerek deneyler yapıldı. Her deney 10 defa tekrarlandı. Yapılan deneylerin ayrıntılı sonuçlarından ziyade ortalama sonuçlarına yer verildi. Yapılan deneylerde kısmi pozitiflerin eklenmesi yaklaşık olarak 5 adımda tamamlandı. 5 adımdan sonrasında kayda değer miktarlarda eklenen yeni örnek olmadı. Bu sebeple adım sayısı genelde 5 adım ile sınırlı tutuldu.

Eşik değer (ED) varsayılan ayarlarda 0.9 olarak belirlenmişti. Eşik değerin 0.8, 0.7, 0.6 ve 0.5 olduğu durumlar için, her adımda ortalama kaç kısmi pozitif örneğin seçilip eğitim kümesine eklendiği Çizelge 4.19'de gösterildi.

Bu sonuçlara göre, olması gerektiği gibi, eşik değer küçültüldükçe daha fazla KP örnek eğitim kümesine seçilir. Bu değerlerden eğitim kümesinin hangi adımda ne kadar genişlediği de çıkarılabilir. Bu deneylerde elde edilen sonuçların ortalaması Çizelge 4.20'de gösterildiği gibidir. KP örneklerin pozitif sınıfa daha yakın olduğu belirtilmişti. Kısmi pozitif örnekler yerine negatif sınıftan alınan örneklem adım işlevine verildi. Her adımda sırasıyla ortalama 19, 6, 3, 1, 2 örnek 0.9 eşik değerini geçebildi. Aynı şekilde gerçek kısmi pozitifler verildiğinde sırasıyla ortalama 119, 61,

Çizelge 4.18: KP örneklerin eğitim kümesine adım-adım eklenmesi deneyinde elde edilen sonuçların ortalaması.

Adımlar	AUC(m)	AUC(s)	PRBE(m)	PRBE(s)	MAP(m)	MAP(s)
-	0.921	0.002	0.235	0.020	0.149	0.007
1	0.920	0.003	0.219	0.015	0.134	0.007
2	0.918	0.003	0.216	0.019	0.133	0.006
3	0.919	0.003	0.216	0.018	0.132	0.006
4	0.919	0.003	0.218	0.017	0.132	0.003
5	0.918	0.004	0.216	0.015	0.131	0.007

Çizelge 4.19: Farklı eşik değerleri ile yapılan deneylerde, her adımda eğitim kümesine eklenen ortalama KP örnek sayısı.

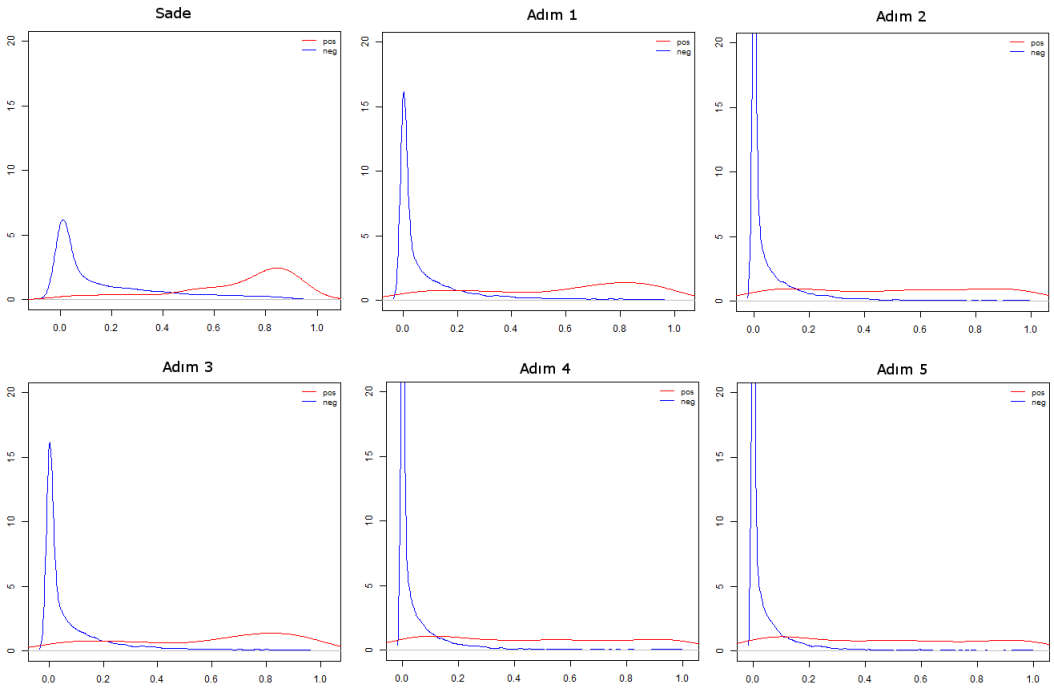
	Eşik Değer				
Adımlar	0.9	0.8	0.7	0.6	0.5
1	119	351	606	802	1018
2	61	96	79	81	104
3	43	28	27	33	45
4	23	15	15	19	26
5	13	10	10	15	15

43, 23, 13 örnek seçilmişti (Çizelge 4.19). Bu sonuç gösteriyor ki kısmi pozitifler, gerçek pozitif sınıfa daha yakın örneklerdir.

Kısmi pozitiflerin eğitim kümesine dahil edilmesi ile ağaçlar büyütülürken kullanılacak pozitif örneklerin sayısı artacağından, her adımda pozitif sınıfın yoğunluğunun artması beklendi (Şekil 4.4). Ancak pozitif sınıfın yoğunluğu artmadı. KP örnekler eklenmeden önce geliştirilen modelde pozitif örnekler, daha çok 0.8 - 0.9 değerleri arasında kümelenirken, KP örneklerin eklenmesiyle daha geniş yayılım gösterdi. Sonuç olarak negatif örnekler negatifliğe yaklaşıp daha çok 0.0 - 0.1 arası değerler alırken, pozitif örnekler yayılıp 0.1 - 0.9 aralığında değerler aldı. RO yöntemindeki örnekleme boyutu parametresi en küçük sınıf boyutuna göre verildiğinden, eğitim kümesine eklenen pozitif örnekler, negatif sınıftan yapılacak örneklemenin de boyutunun artmasına sebep oldu. Eğitim kümesinde pozitif örneklerin yanında negatif örneklerin sayısı da arttı. Bu durumun oluşmasının bir sebebi budur. Bir diğer sebebi de eklenen KP örneklerin gürültülü olmasıdır. RO'da negatif örnekleme sayısı 500'de sabitlendiğinde model, kısmi pozitifleri 0.9 eşik değerine göre ayırmada yetersiz kaldı (her adımda ortalama 3 kısmi pozitif seçilebildi). Bunun için deneylerde, özellikle

Çizelge 4.20: Eşik değer değiştirilerek yapılan deneylerde, her adımda elde edilen sonuçların ortalama değerleri.

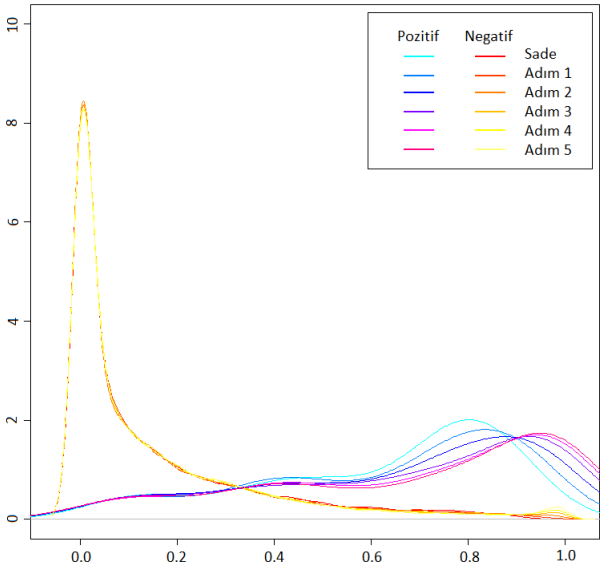
		Eşik Değer									
		0.9		0.8		0.7		0.6		0.5	
Ölçütler	**Adımlar**	Ort.	Sd.	Ort.	Sd.	Ort.	Sd.	Ort.	Sd.	Ort.	Sd.
AUC	1	0.922	0.002	0.920	0.003	0.917	0.003	0.917	0.003	0.914	0.004
	2	0.921	0.002	0.918	0.003	0.917	0.003	0.915	0.004	0.912	0.002
	3	0.922	0.002	0.919	0.003	0.916	0.003	0.914	0.003	0.910	0.003
	4	0.921	0.002	0.919	0.003	0.915	0.004	0.914	0.004	0.910	0.003
	5	0.921	0.001	0.918	0.004	0.915	0.004	0.913	0.003	0.909	0.003
PRBE	1	0.242	0.020	0.219	0.015	0.215	0.010	0.203	0.015	0.214	0.018
	2	0.235	0.016	0.216	0.019	0.221	0.016	0.205	0.019	0.216	0.029
	3	0.233	0.014	0.216	0.018	0.212	0.017	0.205	0.018	0.203	0.018
	4	0.231	0.014	0.218	0.017	0.212	0.019	0.207	0.017	0.210	0.022
	5	0.229	0.012	0.216	0.015	0.208	0.011	0.202	0.010	0.210	0.024
MAP	1	0.154	0.006	0.134	0.007	0.131	0.005	0.124	0.006	0.122	0.009
	2	0.148	0.007	0.133	0.006	0.129	0.009	0.122	0.009	0.118	0.010
	3	0.144	0.008	0.132	0.006	0.124	0.005	0.121	0.007	0.116	0.011
	4	0.144	0.009	0.132	0.003	0.122	0.006	0.118	0.009	0.114	0.009
	5	0.143	0.007	0.131	0.007	0.122	0.005	0.118	0.008	0.115	0.008



Şekil 4.4: RO'da örnekleme boyutu parametresi [k, k] iken her adımda oluşan, pozitif ve negatif örneklerin yoğunluk çizimi. Yoğunluk çiziminde 0-1 arası değişen X ekseni modelden gelen skor değerlerini, y ekseni ise yoğunluğu gösterir. Yoğunluk çiziminin altında kalan alan 1'e eşittir.

pozitif örneklerin sayısı azaltığında, negatif sınıf için 500 yerine daha düşük sayılar kullanıldı. Pozitif örneklerin sayısı yarıya düşürüldüğünde ve örnekleme boyutu

negatif sınıf 100'de sabitlendiğinde Şekil 4.5'deki sonuçlar elde edildi. Bu sonuçlara göre pozitif örneklerin, her adımda pozitife daha çok yaklaştığı görülür. Pozitif için başta 0,8 civarında bir tepe noktası varken, sona doğru bu tepe noktasının 0.9 civarına kaydığı görülür. Negatif örneklerin yoğunluğunun değişmemesinin sebebi her bir ağaç büyütülürken negatif havuzdan sabit olarak 100 rastgele örneklem alınmasıdır. Pozitif



Şekil 4.5: RO'da örnekleme boyutu parametresi [100, k] iken her adımda oluşan, pozitif ve negatif örneklerin yoğunluk çizimi.

örneklerin sayısı düşürüldükten sonra, kısmi pozitiflerin adım adım eklenmesinin model üstünde yapacağı etki test edildi. İlk modelin eğitilmesinde kullanılacak pozitif örneklerin sayısı, yarıya ve çeyreğe düşürülerek iki ayrı deney yapıldı ve Çizelge 4.21'deki ortalama sonuçlar elde edildi. Pozitiflerin %50'sinin kullanıldığı durumda, her adımda sırası ile ortalama 92, 69, 38, 18, 9 KP örnek eğitim kümesine girebildi. Pozitiflerin %25'i kullanıldığında ise sırasıyla ortalama 62, 63, 41, 24, 11 KP örnek eğitim kümesine dahil oldu. Pozitif örneklerin az sayıda kullanılmasının performansı çok fazla düşürmediği daha önce belirtilmişti. Adım sütunu "-" ile gösterilen ilk satırlardaki sonuçlar, KP örnekler kullanılmadan, sade yaklaşımla geliştirilen modele aittir. Bu sonuçlara göre KP örneklerin eklenmesi olumlu ya da olumsuz bir etki göstermedi.

Aynı deney RO'da negatif sınıfın örnekleme boyutu sabitlenerek yapıldı. Negatif örnekleme boyutu pozitif örneklerin yarıya düşürüldüğü durum için 100, çeyreğe düşürüldüğü durum için 50 yapıldı (Çizelge 4.22). Negatif örnekleme sayısı 500

Çizelge 4.21: Pozitif örneklerin sayısının yarıya ve çeyreğe düşürüldüğü durumda, KP örneklerin adım adım eklenmesi ile elde edilen sonuçların ortalaması.

Yüzde (%)	Adımlar	AUC		PRBE		MAP	
		Ort.	Sd.	Ort.	Sd.	Ort.	Sd.
50	-	0.912	0.004	0.200	0.017	0.120	0.007
	1	0.903	0.004	0.214	0.020	0.126	0.010
	2	0.901	0.006	0.216	0.018	0.128	0.010
	3	0.902	0.006	0.218	0.019	0.126	0.009
	4	0.902	0.006	0.218	0.020	0.125	0.008
	5	0.902	0.005	0.210	0.020	0.125	0.011
25	-	0.916	0.003	0.238	0.029	0.146	0.013
	1	0.916	0.003	0.235	0.025	0.143	0.015
	2	0.915	0.002	0.220	0.017	0.134	0.011
	3	0.915	0.003	0.210	0.014	0.129	0.010
	4	0.915	0.003	0.200	0.026	0.125	0.009
	5	0.914	0.003	0.203	0.022	0.124	0.008

seçildiğinde kayda değer örnek seçilmedi. Pozitiflerin sayısı düşürüldüğü ve iki sınıf arasındaki dengesizlik büyüdüğü için bu sonuç elde edildi. Negatif örnekleme boyutu 300 ve 200 yapıldığında da sonuç değişmedi. Bundan dolayı daha küçük değerler kullanıldı. Bu deneyde, pozitiflerin %50'sinin kullanıldığı durumda, her adımda sırası ile ortalama 34, 74, 87, 73, 56 örnek eğitim kümesine eklenmek üzere seçildi. Pozitiflerin %25'i kullanıldığında ise sırasıyla ortalama 19, 59, 109, 128, 101 kısmi pozitif örnek eğitim kümesine dahil edildi.

Adım işlevine KP örneklerden sadece grup-1 olanlar verildiğinde Çizelge 4.23'deki ortalanmış sonuçlar elde edildi. Pozitif örneklerin sayısının düşürüldüğü durumda, adım işlevine KP yerine çıkarılan pozitif örnekler verilerek deneyler yapıldı. Pozitiflerin sayısı yarıya ve çeyreğine düşürüldü, çıkarılan kısım KP olarak adım işlevine verildi (Çizelge 4.24). Bu deneyde eşik değeri 0.9 olduğunda ortalama 16, 5, 3, 2, 2 örnek ayrıldı. Eğitim kümesi çok değişmediği için dikkate değer bir sonuç gözlenmedi. Diğer bir deneyde, adım işlevinde seçim işlemi en yüksek derecelendirilen 30 örnek alınarak yapıldı (Çizelge 4.25). İlk model pozitif örneklerin çeyreği kullanılarak eğitildi. Pozitif örneklerin kalanı adım işlevinde kullanılmak üzere modele verildi. Aynı şekilde eğitilen modele pozitiflerin kalanı yerine KP örnekler verildi. Bu sonuçlara göre pozitif örnekler eğitim kümesine seçilip eklendikçe

Çizelge 4.22: Pozitif örneklerin sayısının yarıya ve çeyreğe düşürüldüğü durumda ve negatif örnekleme sayısı sabitlendiğinde, KP örneklerin adım adım eklenmesi ile elde edilen sonuçların ortalaması. Pozitifler yarıya ve çeyreğe düşürüldüğünde, negatif örneklem sayısı sırası ile 100 ve 50 yapıldı.

Yüzde (%)	Adımlar	AUC		PRBE		MAP	
		Ort.	Sd.	Ort.	Sd.	Ort.	Sd.
50	-	0.916	0.003	0.238	0.029	0.146	0.013
	1	0.916	0.003	0.235	0.025	0.143	0.015
	2	0.915	0.002	0.220	0.017	0.134	0.011
	3	0.915	0.003	0.210	0.014	0.129	0.010
	4	0.915	0.003	0.200	0.026	0.125	0.009
	5	0.914	0.003	0.203	0.022	0.124	0.008
25	-	0.911	0.005	0.233	0.026	0.132	0.010
	1	0.911	0.005	0.219	0.020	0.133	0.011
	2	0.910	0.005	0.212	0.016	0.128	0.010
	3	0.909	0.005	0.204	0.017	0.120	0.006
	4	0.909	0.005	0.188	0.018	0.114	0.004
	5	0.908	0.005	0.185	0.025	0.112	0.006

Çizelge 4.23: Adım işlevine kısmi pozitiflerden sadece grup-1'de olanların verilmesi ile elde edilen sonuçların ortalaması.

Adımlar	AUC(m)	AUC(s)	PRBE(m)	PRBE(s)	MAP(m)	MAP(s)
-	0.922	0.001	0.245	0.018	0.154	0.009
1	0.921	0.001	0.237	0.019	0.150	0.009
2	0.922	0.002	0.229	0.016	0.147	0.009
3	0.921	0.001	0.224	0.015	0.141	0.009
4	0.921	0.002	0.222	0.013	0.139	0.008
5	0.921	0.002	0.216	0.016	0.137	0.005

performans artar. Aynı şekilde kısmi pozitif örnekler eklendiğinde PRBE ve MAP ölçütlerinde ilk üç adımda artış sağlanmıştır. AUC ölçütü ise az miktarda düşmüştür.

Şu ana kadar, farklı yaklaşımlar, bu yaklaşımlarda kullanılan girdi değerleri de değiştirilerek, denendi. İlk olarak, yalnız mevcut altın standart pozitif protein çiftleri kullanıldı ve kısmi pozitifler yok sayıldı. İkinci olarak, bütün kısmi pozitif protein çiftleri, pozitif kabul edilerek, eğitim kümesine eklendi. Böylece kısmi pozitiflerin çok fazla gürültülü olduğu ve pozitif sayılmasının kesinliği düşürdüğü görüldü. Üçüncü olarak, Rastgele Orman metodu beraber öğrenme yapısında, şu şekilde kullanıldı. Pozitif örnekler kullanılarak bir model oluşturuldu. Bu model ile kısmi pozitif örnekler

Çizelge 4.24: Pozitiflerin sayısı yarıya ve çeyreğine düşürülüp, çıkarılan kısım KP gibi adım işlevine verildiğinde elde edilen sonuçların ortalamaları.

Yüzde (%)	Adımlar	AUC		PRBE		MAP	
		Ort.	Sd.	Ort.	Sd.	Ort.	Sd.
2	-	0.923	0.009	0.232	0.031	0.146	0.028
	1	0.923	0.011	0.230	0.036	0.149	0.033
	2	0.923	0.010	0.238	0.041	0.152	0.032
	3	0.923	0.010	0.232	0.036	0.150	0.033
	4	0.923	0.010	0.234	0.032	0.150	0.032
	5	0.923	0.010	0.232	0.039	0.143	0.027
4	-	0.903	0.016	0.196	0.060	0.122	0.025
	1	0.900	0.017	0.212	0.042	0.128	0.024
	2	0.901	0.017	0.217	0.046	0.128	0.025
	3	0.900	0.018	0.228	0.056	0.135	0.035
	4	0.902	0.016	0.224	0.050	0.138	0.026
	5	0.902	0.016	0.225	0.056	0.141	0.035

Çizelge 4.25: Adım işlevinde kullanılan pozitif örneklerin kalanı ile normal KP örneklerin karşılaştırılması.

Eklenen	Adımlar	AUC		PRBE		MAP	
		Ort.	Sd.	Ort.	Sd.	Ort.	Sd.
Kalan Pozitifler	-	0.904	0.018	0.208	0.066	0.135	0.044
	1	0.904	0.018	0.215	0.051	0.141	0.038
	2	0.902	0.018	0.230	0.052	0.152	0.038
	3	0.907	0.018	0.240	0.070	0.171	0.054
	4	0.916	0.017	0.234	0.055	0.165	0.049
	5	0.923	0.010	0.232	0.039	0.143	0.027
Kısmi Pozitifler	-	0.915	0.016	0.200	0.036	0.150	0.045
	1	0.913	0.017	0.240	0.043	0.160	0.038
	2	0.911	0.017	0.254	0.046	0.170	0.037
	3	0.911	0.020	0.249	0.045	0.164	0.037
	4	0.910	0.020	0.267	0.063	0.164	0.034
	5	0.902	0.016	0.225	0.056	0.141	0.035

sınıflandırdı. Bu sınıflandırma işleminin sonucunda yüksek değerde sınıflandırılan örnekler eğitim kümesine eklenerek model yeniden eğitildi. Bu işlem eğitim kümesine eklenecek örnek kalmayıncaya ya da eklenecek örnek sayısı önemsiz düzeye gelinceye kadar devam ettirildi. Fazla gürültü barındırmalarından dolayı kısmi pozitiflerin bütünüyle pozitif sayılması performansı olumsuz yönde etkiledi. Beraber öğrenme yapısında modele dahil edilmeleri, performans üzerindeki olumsuz etkilerini düşürdü. Bizim öngörümüz ek girdiyi bu yöntem ile kullanarak model performansı arttırmaktı.

Ancak model performansı istenilen ölçüde artmadı. Pozitif örneklerin niteliği bunun başlıca sebebi olarak yorumlandı. Pozitif örneklerin yarısı kullanılarak geliştirilen modelin performansı, tamamının kullanıldığı duruma göre pek farklılık göstermedi. Bu sonuç, pozitif örneklerin birbirine benzediği ve insan-HIV arasındaki etkileşim kümesinin tamamını temsil edecek şekilde yeterince kapsayıcı olmadıkları fikrini verdi.

5. SONUÇ VE ÖNERİLER

Gözetimli yapay öğrenme yöntemleri modeli eğitebilmek için yeterli sayıda etiketli veriye ihtiyaç duyar. Ancak çok çalışılmış olanlar haricinde, organizmalar arası protein-protein etkileşimini veren yeterince büyük veri kümeleri pek bulunmaz. Etiketli örneklerin az olması, geleneksel yapay öğrenme yöntemleri dışında farklı stratejiler kullanmayı gerekli kılar. Bunlardan biri de yarı gözetimli öğrenme yöntemleri ile ek bilgiler kullanarak, modelin gücünü arttırmaktır.

Bu çalışmada HIV-1 ve insana ait protein-protein çiftlerinin etkileşip etkileşimediğinin tahmin edilmesinde gözetimli yapay öğrenme yöntemleri kullanıldı. Ek girdi olarak, kısmi pozitif protein çiftlerinin daha etkili bir şekilde kullanımına yönelik farklı yaklaşımlar geliştirildi. Kısmi pozitifler literatürde farklı anahtar kelimelerle birlikte geçen ve pozitife yakın olan protein çiftleridir.

Ek girdiyi kullanmak için farklı yaklaşımlar denenmiştir. Sade yaklaşımda, mevcut pozitif protein çiftleri ile negatif kümeden örneklenen protein çiftleri kullanılmıştır. Bu yolla, kısmi pozitiflerin yok sayıldığı durumdaki başarı ölçüldü. İkinci bir yaklaşım olarak bütün kısmi pozitif protein çiftleri, pozitif varsayılarak, eğitim kümesine eklendi. Böylece kısmi pozitiflerin çok fazla gürültülü olduğu ve pozitif sayılmasının kesinliği düşürdüğü görüldü. Önerilen yaklaşım ise, kısmi pozitiflerin eklenmeden önce teste tabi tutulmasıdır. Bu yaklaşımda mevcut bilinen pozitif protein çiftleri ve örneklenen negatif protein çiftleri ile ilk model geliştirilir. Geliştirilen bu model ile kısmi pozitif örnekler test edilir. Ancak belirli bir eşik değeri geçen örnekler eğitim kümesine dahil edilir. Ekleme işleminden sonra oluşan genişletilmiş eğitim kümesi ile model yeniden geliştirilir. Bu işlem kayda eklenecek anlamlı miktarda örnek kalmayıncaya kadar devam ettirilir. Bu yaklaşım yapısı itibari ile birlikte öğrenmeye yöntemine benzer.

Sonuç olarak en iyi performans kısmi pozitif örneklerin yoksayıldığı sade yaklaşımda elde edildi. Bütün kısmi pozitif örneklerin doğru kabul edilerek eğitim kümesine dahil edilmesi geliştirilen modelin performansını olumsuz yönde etkiledi. Öte yandan

kısmi pozitif örneklerin birlikte öğrenme yapısında adım adım eğitim kümesine dahil edilmesi, tamamın doğru kabul edilerek model geliştirilirken kullanıldığı yaklaşıma göre daha iyi sonuç verdi. Ayrıca bu yaklaşım ile kısmi pozitif örneklerin kullanılmasından doğan kesinlik değerindeki düşüşün de önüne geçildi. Ancak performans öngörüldüğü biçimde arttırılamadı. Pozitif örneklerin niteliği bunun başlıca sebebi olarak yorumlandı. Pozitif örneklerin yarısı kullanılarak geliştirilen modelin performansı, tamamının kullanıldığı duruma göre pek farklılık göstermedi. Bu sonuç, pozitif örneklerin birbirine benzediği ve insan-HIV arasındaki etkileşim kümesinin tamamını temsil edecek şekilde yeterince kapsayıcı olmadıkları fikrini verdi.

KAYNAKLAR

[1] **Gonzalez, M.W. ve Kann, M.G.** (2012). Chapter 4: Protein Interactions and Disease, *PLoS Comput Biol*, ***8***(12), e1002819.

[2] **Rivas, J.D.L. ve Fontanillo, C.** (2010). Protein-Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks., *PLoS Computational Biology*, ***6***(6).

[3] **Trkola, A.**, (2004), HIV-host interactions: vital to the virus and key to its inhibition, Curr Opin Microbiol. 7, 555-9.

[4] **Qi, Y.**, **Tastan, O.**, **Carbonell, J.**, **Klein-Seetharaman, J. ve Weston, J.** (2010). Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins, *Bioinformatics*, ***26***(18), i645.

[5] **Frankel AD, Y.J.** (1998). HIV-1: fifteen proteins and an RNA., *Annu Rev Biochem*, ***67:1-25***.

[6] **Özlem Aker**, (2010), "HIV Virüsü", `http://www.duzen.com.tr/eJournals/2010/Bulten-Kasim2010.pdf`.

[7] **Babayiğit, M. A. ve Bakır, B.** (2004). HIV enfeksiyonu ve AIDS: epidemiyoloji ve korunma, *TAF Prev Med Bull.*, ***63***.

[8] **Mohri H., Perelson AS., T.K.v.d.** (2001). Increased turnover of T lymphocytes in HIV-1 infection and its reduction by antiretroviral therapy., *J Exp Med*, ***194(9):1277-87.***

[9] **B. Ahr, V. Robert-Hebmann, C.D. ve Biard-Piechaczyk, M.** (2004). Apoptosis of uninfected cells induced by HIV envelope glycoproteins, *Retrovirology*, ***10.1186/1742-4690-1-12***.

[10] **WHO, UNICEF, U.**, (2013), Global update on hiv treatment: results, impact and opportunities, `http://www.who.int/hiv/pub/progressreports/update2013/en/`.

[11] **Aktürkoğlu, E.**, (2012), HIV and AIDS estimates of Turkey, `http://www.unaids.org/en/regionscountries/countries/turkey/`, alındığı tarih: 22.11.2013.

[12] **Tastan, O.**, **Qi, Y.**, **Carbonell, J.G. ve Klein-Seetharaman, J.** (2009). Prediction of Interactions Between HIV-1 and Human Proteins by Information Integration., ***R.B. Altman, A.K. Dunker, L. Hunter, T. Murray ve T.E. Klein****, (düzenleyenler), Pacific Symposium on Biocomputing*.

[13] **Qi, Y.**, **Bar-joseph, Z. ve Klein-seetharaman, J.** (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction, *Proteins*, ***63***.

[14] **Oznur Tastan, Yanjun Qi, J.G.C.v.J.K.S.**, Supporting online material for prediction of interactions between hiv-1 and human proteins by information integration, `http://www.cs.cmu.edu/~oznur/hiv/hivPPI.html`, alındığı tarih: 07.06.2012.

[15] **Mohamed, T.P.**, **Carbonell, J.G. ve Ganapathiraju, M.** (2010). Active learning for human protein-protein interaction prediction., *BMC Bioinformatics*, ***11***(S-1), 57.

[16] **Yip, K.Y. ve Gerstein, M.** (2009). Training Set Expansion: An Approach to Improving the Reconstruction of Biological Networks from Limited and Uneven Reliable Interactions, *Bioinformatics*, ***25***(2), 243–250.

[17] **Shi, M. ve Zhang, B.** (2011). Semi-supervised learning improves gene expression-based prediction of cancer recurrence, *Bioinformatics*, ***27***(21), 3017–3023.

[18] **Wang, X. ve Simon, R.** (2011). Microarray-based cancer prediction using single genes, *BMC Bioinformatics*, ***12***(1), 391.

[19] **Chapelle, O. ve Zien, A.** (2005). Semi-supervised classification by low density separation, *Proceedings of the International Workshop on Artificial Intelligence and Statistics*.

[20] **Y. Qi, O. Tastan, J.C.J.K.S.v.J.W.**, Supporting online material for semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins, `http://www.cs.cmu.edu/~qyj/HIVsemi/`, alındığı tarih: 07.06.2012.

[21] **Fu, W.**, **Sanders-Beer, B.E.**, **Katz, K.S.**, **Maglott, D.R.**, **Pruitt, K.D. ve Ptak, R.G.** (2009). Human immunodeficiency virus type 1, human protein interaction database at NCBI, *Nucl. Acids Res.*, ***37***.

[22] **Puntervoll, P.**, **Linding, R. ve Gemünd, v.d.** (2003). ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins., *Nucleic Acids Research*, ***31***(13).

[23] **Wang, J.Z.**, **Du, Z.**, **Payattakool, R.**, **Yu, P.S. ve Chen, C.F.** (2007). A New Method to Measure the Semantic Similarity of GO Terms., *Bioinformatics*.

[24] **Alpaydin, E.** (2010). *Introduction to Machine Learning*, The MIT Press.

[25] **Breiman, L.** (2001). Random Forests, *Mach. Learn.*, ***45***(1), 5–32.

[26] **Christopher D. Manning, Prabhakar Raghavan, H.S.** (2008). *Introduction to Information Retrieval*, Cambridge University Press.

[27] **Liaw, A. ve M. Wiener, F.o.b.L.B.**, (2012), Package 'randomForest', `http://cran.r-project.org/web/packages/randomForest/index.html`, alındığı tarih: 02.12.2013.

[28] **Url-1**, http://www.unaids.org/en/dataanalysis/datatools/aidsinfo/, alındığı tarih: 22.11.2013.

[29] **Url-2**, http://en.wikipedia.org/wiki/HIV, alındığı tarih: 20.11.2013.

[30] **Url-3**, http://tr.wikipedia.org/wiki/Apoptozis, alındığı tarih: 20.11.2013.

[31] **Url-4**, www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#workings, alındığı tarih: 15.12.2013.

[32] **Url-5**, http://www.cs.waikato.ac.nz/ml/weka/, alındığı tarih: 02.12.2013.

[33] **Url-6**, http://www.r-project.org/, alındığı tarih: 02.12.2013.

[34] **Url-7**, http://en.wikipedia.org/wiki/Random_forest, alındığı tarih: 15.01.2014.

EKLER

EK A. Sözlük

EK A

Türkçe - İngilizce	
Algılayıcı	Perceptron
Altın standart	Gold-standard
Anma	Recall
Ayrım	Discrimination
Bağlanım	Regression
Birlikte öğrenme	Co-training
Box plot	Kutu çizgesi
Çapraz düzensizlik	Cross-entropy
Çapraz sağlama	Cross-validation
Çevrimiçi güncelleme kuralı	Online update rule
Çıkıt birimi	Output unit
Çok katmanlı algılayıcı	Multilayer Perceptron
Çoklu görev	Multi-task
Değişke	Variance
Density plot	Yoğunluk çizgesi
Doğru pozitif	True positive
Doğruluk	Accuracy
Düğüm	Node
Düzensizlik	Entropy
Eğim iniş	Gradient descent
Ek girdi	Bias unit
Entropi	Düzensizlik
Eşiksiz en büyük işlev	Softmax (function)
Fold	Kat
Gen düzenleyici ağı	Gene regulatory network
Gen ifadesi belirleme	Gene expression profiling
Girdi birimi	Input unit
Gözetimli	Supervised
Gözetimli öğrenme mimarisi	Supervised learning framework
Ham	Raw
Hassaslık	Sensitivity
Hata işlevi	Error function
İç çarpım	Dot product
Karar ağacı	Decision tree
Kesinlik	Precision
Kısmi pozitif	Partially positive
Kısmi sınıflandırılmış	Partially labeled
Örnekleme	Sampling
Özgüllük	Specificity
Proteaz enzimi	Protease enzyme
Rastgele orman	Random forest
S işlemi	Sigmoid function
Saflık ölçütü	Impurity measure

Saklı birim	Hidden unit
Saklı katman	Hidden layer
Sapma	Bias
Sonsal	Posterior
Tahmin yayılımı	Prediction propagation
Ters transkriptaz	Reverse transcriptase
Torba-dışı	Out-of-bag
Yanlış pozitif	False positive
Yapay sinir ağları	Artificial neural network
Yarı gözetimli	Semi-supervised
Zincir kuralı	Chain rule

ÖZGEÇMİŞ

Ad Soyad: İsmail BİLGEN

Doğum Yeri ve Tarihi: Siirt - 23.03.1988

E-Posta: ibilgen@itu.edu.tr

Lisans: Marmara Üniversitesi (2010)

Y. Lisans: İstanbul Teknik Üniversitesi (2014)

TEZDEN TÜRETİLEN YAYINLAR/SUNUMLAR

- **Bilgen, İ.**, Saraç, Ö., Özgür, A., Çataltepe, Z., Co-training using Random Forests for Predicting Human-HIV Protein Interactions, *International Symposium on Health Informatics and Bioinformatics(HIBIT), 2013 8th* .