TÜBİTAK

Zenginleştirilmiş Öznitelikler ve Makine Öğrenmesi Yöntemleriyle Protein Yerel Yapı Tahmini

Program Kodu: 3501

Proje No: 113E550

Proje Yürütücüsü:

Yrd. Doç. Dr. Zafer AYDIN

AĞUSTOS 2017

KAYSERİ

ÖNSÖZ

Protein yapı tahmini biyoenformatik ve teorik kimyanın en önemli hedeflerinden biridir. Deneysel yöntemlerin yetersiz kaldığı durumlarda protein yapısının hesaplama yöntemleri ile tahmin edilmesi etkili ve verimli bir yaklaşımdır. Buna ek olarak proteinin yapısı ile işlevi arasında yakın bir ilişki olduğundan yapının doğu tahmin edilmesi proteinin fonksiyonu için önemli ipuçları verir. Ayrıca tıpta ilaç tasarımı ve biyoteknolojide yeni enzimlerin tasarlanması gibi uygulamaları da olduğundan protein yapı tahmini yüksek derecede önem taşımaktadır. Bu projede protein yapı tahmininde kullanılan ikincil yapı, dihedral açı ve çözücü erişilirlik gibi bir boyutlu yapısal özelliklerin tahmin edilmesi ve parçacık seçimi için öznitelik çıkarma, boyut düşürme, derin öğrenme ve topluluk yöntemleri geliştirilmiştir.

İÇİNDEKİLER

ÖZET

Projenin amacı proteinlerde bulunan ikincil yapı, dihedral açı ve çözücü erişilirlik gibi bir boyutlu yapısal özelliklerin başarılı olarak tahmin edilmesi ve bu tahminleri kullanarak parçacık seçimi yapan yeni bir yöntem geliştirilmesidir. Geliştirilen yöntemler sayesinde proteinlerin üç boyutlu yapısının daha doğru tahmin edilmesi, proteinlerin fonksiyonlarının daha iyi anlaşılması ve daha etkili ilaç tasarımı yapılması mümkün olacaktır. Bir boyutlu yapısal özelliklerin tahmini için yürütücünün daha önce geliştirdiği iki aşamalı hibrit sınıflandırma yöntemi kullanılmıştır. Bu yöntemde bulunan sınıflandırıcılar için dizi tabanlı

profiller, yapısal profil matrisleri gibi çeşitli öznitelik vektörleri kullanılmıştır. İkinci aşamadaki sınıflandırıcı için destek vektör makinası, derin KSA, rastgele orman ve topluluk gibi çeşitli öğrenme yöntemleri eğitilmiş ve geliştirilen yöntemlerin tahmin başarı oranları standart veri kümelerinde incelenmiştir. Ayrıca bu aşamada derin otokodlayıcılar ve öznitelik seçme yaklaşımları ile boyut düşürme gerçekleştirilmiştir. Protein parçacık seçimi için verilen iki amino asit dizisi parçacığının yapısal olarak benzer olup olmadığının tahmin eden yöntemler geliştirilmiştir. Bunun için Rosetta programının parçacık veritabanında bulunan proteinlerden parçacık ikilileri örneklenmiş, bu ikililer BCScore yöntemi ile etiketlenmiş, eğitim ve test kümeleri oluşturulmuştur. Ayrıca farklı öznitelik kümeleri konsept hiyerarşi yaklaşımı ile kapsamlı olarak incelenmiş ve en başarılı sonucu veren öznitelik kombinasyonları tespit edilmiştir. Parçacık seçimi probleminde 3 ve 9 amino asitlik parçacıklar üzerinde çalışılmıştır ancak yöntemler diğer uzunluktaki parçacıklar için de kolaylıkla uygulanabilecektir. Projede geliştirilen yöntemler sayesinde ikincil yapı tahmin başarısı en zor tahmin kategorisinde %2.6 iyileşmiş, dihedral açı tahmin başarısı önemli oranda iyileşmiş, çözücü erişilirlik probleminde literatürdeki en başarılı yöntemler ile benzer bir seviye yakalanmıştır. Parçacık seçiminde ise verilen iki parçacığın yapılarının benzer olup olmadıkları 3-mer parçacıklar için %94 ve 9-merler içinse %97 oranı ile tahmin edilmiştir. Yapılan çalışmaların neticesinde öznitelik vektörlerinin daha iyi tasarlanmasının ve farklı sınıflandırma yöntemlerinin birleştirilip optimize edilmesinin yapısal özellik tahmin başarısını önemli oranda iyileştirdiği sonucuna varılmıştır.

Anahtar kelimeler: Bir boyutlu protein yapı tahmini, protein parçacık seçimi, makine öğrenmesi, derin öğrenme, öznitelik çıkarımı, boyut düşürme

ABSTRACT

The current project concentrated on predicting one dimensional structural properties of proteins such as secondary structure, dihedral angle and solvent accessibility successfully and developing a novel method that uses these predictions for fragment selection. Upon reaching these objectives it is anticipated that the accuracy and quality of protein 3D structure prediction will improve, which will provide a better understanding of the functional roles of proteins and advance drug screening, drug design, and enzyme design processes.

To predict one dimensional structural properties a two-stage hybrid method is used, which employs sequence based profiles and structural profiles as input features. For the classifier at the second stage support vector machine, deep CNF, random forest and an ensemble classifier have been trained and tested on established benchmarks. Additionally, dimensionality reduction techniques are developed and analyzed at this stage including deep autoencoders and feature selection methods. For fragment selection, classifiers have been developed that decide whether two amino acid fragments are structurally similar or not. To build the train and test sets, fragment pairs are sampled from the fragment database of the Rosetta program and labeled using BCScore method. A concept hierarchy approach has been implemented to find the best feature set combination. Though the present study concentrated on 3-mers and 9-mers the methods developed can also be applied easily to other fragment sizes. According to evaluations, a 2.6% improvement has been obtained for protein secondary structure prediciton in the most difficult setting, a significant improvement in dihedral angle class prediction, and an accuracy comparable to state-of-the-art methods in solvent accessilibity. In fragment selection fragment pairs can be classified as similar or not with 94% accuracy for 3-mers and 97% for 9-mers. As a result, designing better features, combining and optimizing classifiers improve the success rates of methods that predict structural properties of proteins.

Keywords: one dimensional protein structure prediction, protein fragment selection, machine learning, deep learning, feature extraction, dimensionality reduction

1. GİRİŞ

İnsan Genom Projesi gibi büyük çaplı DNA dizileme çalışmalarından çok fazla sayıda protein dizi verisi üretilmektedir. Ancak bu proteinlerin birçoğunun yapısı deneysel olarak çözülmemiştir. Gelinen noktada üç boyutlu yapısı deneysel olarak çözülenler dizisi bilinen proteinlerin %0.6'dan daha azını oluşturmaktadır. X-ışını kristalografisi ve Nükleer Manyetik Rezonans (NMR) gibi protein yapısını deneysel olarak bulan yöntemler yoğun işgücü gerektirebilmekte, zaman almakta ve masraflı olmaktadır. Ayrıca bazı protein yapılarının (örn. bazı membran proteinleri) deneysel olarak bulunması mümkün olmamaktadır. Deneysel yöntemlerin yetersiz kaldığı durumlarda protein yapısının hesaplama yöntemleri ile tahmin edilmesi etkili ve verimli bir yaklaşımdır. Diğer yandan protein yapısı ile işlevi arasında yakın bir ilişki olduğundan bir proteinin yapısının bilinmesi onun biyolojik işlevi ve moleküler mekanizmasının aydınlatılması hakkında önemli ipuçları vermektedir. Tıpta ilaç tasarımı ve biyoteknolojide yeni enzimlerin tasarlanması gibi uygulamaları da olduğundan yapı tahmini yüksek derecede önem taşımaktadır.

Protein yapı tahmini, amino asit dizisi verilen bir proteinin üç boyutlu yapısının hesaplanması olarak tanımlanabilir. Bunun için protein molekülündeki bütün atomların üç boyutlu uzaydaki koordinatlarının belirlenmesi gerekmektedir. Günümüze dek protein yapı tahmini üzerinde çok sayıda araştırma yapılmıştır. Buna rağmen yapı tahmin problemi henüz tam olarak çözülememiştir. Üç boyutlu yapı tahmini oldukça zor bir problemdir. Yapının doğrudan doğruya tahmin edilmesinin çeşitli güçlükleri olduğundan aşama aşama ilerlenir. Örneğin önce hedef protein veritabanındaki proteinlerle çeşitli hizalama algoritmalarıyla kıyaslanır ve amino asitlerin belirli pozisyonlarda görülme sıklıklarını özetleyen istatistiksel profil matrisleri oluşturulur. Daha sonra bu matrisler kullanılarak protein yapısının çeşitli özellikleri tahmin edilir. Bunlar arasında ikincil yapı, dihedral açılar, çözücü erişilirlik, bağlanma bölgesi, fonksiyonel bölge, etki alanı sınırı, düzensiz bölge gibi bir boyutlu (sembol dizisi ile gösterilen) yapısal özelliklerin ve yakınlık haritası gibi iki boyutlu (matris ile gösterilen) özelliklerin tahmin edilmesi (öngörülmesi) sayılabilir. Bir sonraki aşamada ise bu özellikler profil matrisleriyle ve diğer fiziksel prensipler ile birlikte kullanılarak proteinin üç boyutlu yapısı tahmin edilir. Her iki senede bir mevcut yapı tahmini yöntemlerinin başarısı dünya genelinde yapılan CASP yarışması ile ölçülmektedir.

Protein yapı tahmini ilaç tasarımı çalışmalarında da kullanılmaktadır. Protein-ligand etkileşimlerinin doğru tanımlanması moleküler biyoloji ve farmakoloji alanlarındaki çalışmalar için önemlidir. Bu amaca yönelik olarak ligand-bağlı reseptör komplekslerinin yapıları X-ışını

kristalografisi ve NMR ile, bağlanma enerjileri hız sabitlerinden ve bağlanmada önemli olan amino asitler mutagenez çalışmaları sonucunda tanımlanmıştır. Bu deneyler her ne kadar protein-ligand kompleksinin yeterli düzeyde tanımlanmasına katkı sağlasa da genellikle büyük bir çaba gerektirir ve rutin olarak gerçekleştirilmeleri zordur. Protein-ligand komleksleri ile ilgili benzer bilgiler, moleküler kenetlenme ve moleküler dinamik (MD) simülasyonları gibi moleküler modelleme teknikleri kullanılarak daha kolay elde edilebilir. MD simülasyon teknikleri deneysel verilerin bulunmadığı durumlarda dahi detaylı dinamik bilgi sağlayabilmektedir. Örneğin ilaç geliştirme aşamalarında standart veya yeni geliştirilmiş bir molekülün spesifik bir proteine nasıl bağlandığı, bu moleküllerin bağlanma sonrasında reseptör yapısı üzerine olan etkileri, ligand-reseptör etkileşimi sırasında gözlenen yapısal değişiklikler ve ligand ve proteinler arasındaki temel etkileşimlerin aydınlatılması gibi ilaç geliştirme alanında önemli sayılan konularda bilgi sağlayabilir. Bu simülasyonlarda proteinin üç boyutlu yapı bilgisi için deneysel veriler kullanılabildiği gibi tahmin edilen koordinat değerleri de kullanılabilmektedir. Ayrıca MD simülasyonları sayesinde proteinlerin tahmin edilen yapılarının iyileştirilmesi de mümkün olmaktadır. Ligand-protein etkileşimlerinin tanımlanması, ilaç tasarımı ve devamında yeni tedavi modellerinin araştırılmasında kritik öneme sahiptir.

2. LİTERATÜR ÖZETİ

Hesaplamalı tekniklerin kullanımı, biyolojik verilerin üssel büyümesi, karmaşıklığı ve erişilebilirliği nedeniyle biyoinformatikte yoğun bir ilgi görmüştür. Bu büyük miktarda veride gizlenen bilgiyi keşfetmeye doğru giden yolda, makine öğrenme yaklaşımları önemli bir rol oynamaktadır. Makine öğrenme yaklaşımları; protein yapı tahmini (Jones, 2001), protein dizi analizi (Zeng vd., 2009), protein katman tanımlama (Bologna ve Appel, 2002; Chinnasamy vd., 2005; Ding and Dubchak, 2001), protein fonksiyon tahmini (Bhola vd. 2014), gen ağı çıkarımı (Perrin vd., 2003), metabolik yol analizi (Dale, 2010) gibi birçok probleme başarıyla uygulanmıştır.

Protein ikincil yapısı, protein yapısını stabilize eden düzenli hidrojen bağlanma örüntüleriyle oluşur (Kabsch ve Sander, 1983). Protein ikincil yapı tahmini (PSSP) heliks (helix), iplik (strand) ve döngüyü (loop) içeren üç harfli (H, E, L) alfabeden yapısal bir durum atamayı amaçlamaktadır. İkincil yapıyı tahmin etmek için, tipik olarak bir modelin Protein Data Bank (PDB) 'de bulunan üç boyutlu yapı bilgilerinden türetilen ikincil yapı etiketleri bilinen proteinleri kullanarak eğitildiği denetimli öğrenme kullanılır (Bernstein vd., 1977).

Bir boyutlu yapısal özelliklerin tahmini için çeşitli yöntemler önerilmiştir. Destek vektör

makineleri (SVM) ve yapay sinir ağları, diğer yöntemlere kıyasla daha iyi doğruluk oranları vermektedir. Destek vektör makineleri kullanılarak çeşitli sınıflandırıcılar geliştirilmiştir (Ward vd., 2003; Hua ve Sun, 2001; Kim ve Park, 2003; Gubbi vd., 2006; Chen vd., 2008; Huang ve Chen, 2013; Wang vd., 2016). Yapay sinir ağları kullanılarak yapılan çalışmalarda da olumlu sonuçlar alınmıştır (Jones 1999; Pollastri vd., 2002b; Pollastri ve McLysaght 2005; Mirabello ve Pollastri 2013; Jian-wei vd., 2013). Ayrıca yakın zamanda derin öğrenme yaklaşımları ile başarılı sonuçlar alınmıştır (Spencer vd., 2015; Li ve Yu 2016; Wang vd., 2016). Bunlara ek olarak saklı Markov modelleri (Martin vd., 2006; Aydin vd., 2006), dinamik Bayes ağları (Yao vd., 2008; Aydin vd., 2008), en yakın komşu (Salamov ve Solovyev, 1995; Ghosh 2008; Yang vd., 2011) gibi yöntemler de kullanılmıştır.

Topluluk öğrenimi model tanımlama, makine öğrenimi ve veri madenciliğinde önemli bir tekniktir. Topluluk öğreniminin arkasındaki temel fikir, doğruluk oranını artırmak için birden fazla sınıflandırıcıyı birleştirmektir (Dietterich, 2000). Son zamanlarda, farklı yöntemleri birleştirerek doğruluğu geliştirmek için birçok çalışma yapılmıştır (King vd., 2000; Kountouris vd., 2012; Alirezaee vd., 2012; Pollastri vd., 2002a). Toplulukların yanı sıra, çeşitli sınıflandırıcıların güçlü yönlerini birleştiren hibrit yöntemler de vardır (Yao vd., 2008; Aydın vd., 2011; Wang vd., 2016).

Bir boyutlu yapısal özelliklerin tahmin başarısını iyileştirmek için sınıflandırıcıların iyi tasarlanmasına ek olarak etkili özniteliklerin çıkarımı da oldukça önemlidir. Bu alanda daha ziyade PSI-BLAST (Altschul vd., 1997) ve HHBlits (Remmert vd., 2012) gibi yöntemler ile hesaplanan özniteliklere ek olarak yapısal profil matrisleri de kullanılmaktadır (Li vd., 2011). Bunlar arasında en yaygın olarak PSI-BLAST ile hesaplanan profil matrisleri kullanılmıştır. HHBlits ile hesaplanan profil matrisleri ile yapısal profil matrislerini kullanan çalışma sayısı nispeten daha azdır. Ayrıca yapısal profil matrislerinin hesaplanması için basit dizi hizalama teknikleri ve sadece belirli hizalama skorlarına bağlı basit ağrılık katsayısı modelleri kullanılmaktadır. Bu katsayıların daha iyi modellenmesinin tahmin başarı oranını iyileştirdiği yürütücü tarafından gösterilmiştir.

Proteinlerin yapısal özelliklerini tahmin etmekte karşılaşılan problemlerden birisi de yeni öznitelik çıkarma yaklaşımlarının geliştirilmesi ile veri uzayının boyutunun artmasıdır. Bu durumda model eğitme zamanları özellikle destek vektör makinası gibi yöntemler için önemli oranda artmaktadır. Ayrıca aşırı uyum davranışına takılma riski de artmakta ve bunun sonucunda tahmin başarı oranları düşmektedir. Çözüm olarak izlenecek yaklaşımlardan birisi boyut düşürmedir. İkincil yapı tahmini için çeşitli boyut düşürme yöntemleri kullanılmıştır (Adamczak, 2009; Li vd., 2017). Ancak boyut düşürme yaklaşımları kullanılarak proteinlerin

yapısal özelliklerinin tahmin eden sınırlı sayıda çalışma vardır.

Projede bir boyutlu yapısal özelliklerin tahminine ek olarak parçacık seçimi problemi üzerinde çalışılmıştır. Parçacık seçimi üç boyutlu yapı tahmini yapan yöntemlerde yaygın olarak kullanılmaktadır. Özellikle kalıp proteinlerin bulunamadığı serbest modelleme yöntemlerinde (Fujitsuka vd., 2006; Gront vd., 2011; Lee vd., 2004, 2011; Li vd., 2008; Simoncini vd., 2012; Tian vd., 2011; Xu ve Zhang, 2012) ve hedef proteinin kalıplarla eşleşemeyen bazı alt bölgelerini modellemek için (Cheng vd., 2012; Lee vd., 2010; Roy vd., 2010; Zhang vd., 2012) tercih edilmektedir. Parçacık seçimi için geliştirilen yöntemler genellikle logistic regresyon gibi basit doğrusal modeller kullanmaktadır. Doğrusal olmayan modellerin kullanıldığı çalışmalar sınırlı sayıdadır. Doğrusal modellerin ağırlık parametreleri daha ziyade makine öğrenmesi yaklaşımı ile (eğitim test kümeleri kullanılarak) değil, deneme yanılma yaklaşımı ile sınırlı bir aralıkta belirli değerlerin seçilip yapı tahmini başarı oranlarını iyi veren kombinasyonların tespit edilmesi ile bulunmaktadır. Buna ek olarak parçacıkların benzerliği için daha ziyade RMSD skoru kullanılmıştır. Parçacık yapılarının benzerliğini daha iyi modelleyen BCScore metriği ise henüz kullanılmamıştır. Ayrıca en uygun öznitelik kombinasyonunu konsept hiyerarşisi yaklaşımı ile kapsamlı ve sistematik olarak araştıran bir çalışma da bulunmamaktadır.

3. GEREÇ VE YÖNTEM

3.1 Öznitelik Çıkarma

Bu tezde, iki tür öznitelik kullanılmıştır: sıralı hizalamalardan türetilmiş pozisyona özgü puanlama matrisleri (PSSM) ve yapısal profil matrisleri. PSSM'ler, PSI-BLAST (Altschul vd., 1997) ve HHBlits (Remmert vd., 2012) programları ile hesaplanmıştır ve PSI-BLAST PSSM ve HHMAKE PSSM olarak adlandırılır. Bunlar Dinamik Bayes Ağları (DBN) ve hibrid yöntemin (DSPRED) ikinci aşamasındaki sınıflandırıcılarda öznitelik olarak kullanılır. Yapısal profil matrisleri ise HHBlits programının ikinci aşamasından sonra hedef proteine (target) hizalanmış PDB proteinlerinin yapı etiketlerini kullanarak hesaplanmaktadır.

3.1.1 PSIBLAST PSSM

PSI-BLAST yöntemi BLAST algoritmasının yinelemeli versiyonu olarak düşünülebilir (Altschul vd., 1997). Verilen protein dizisi veritabanındaki proteinlerle ikili hizalanarak eşik değerinin üzerinde kalan proteinler seçilir. İkinci ve daha sonraki yinelemelerde ise eşik değerinin üzerinde kalan proteinler çoklu hizalama yöntemiyle hizalanır ve bir istatistiksel profil matrisi hesaplanarak veritabanındaki protein dizileriyle hizalanır. Bu hizalama sonucunda eşik değerinin üzerinde kalan proteinler bir sonraki yineleme için saklanır ve her yinelemede eşik

değerinin üzerinde kalan proteinler kullanılarak profil matrisi güncellenir. Genellikle 3-6 yineleme, sonuçların yakınsaması için yeterli olmaktadır. PSI-BLAST yöntemiyle elde edilen profil matrisinin boyutu 20×U'dir öyle ki U hedef proteindeki amino asit sayısıdır. Hizalamada profillerin kullanılmasıyla dizisel benzerliği az olup yapısal benzerliği olan proteinler de keşfedilebilmekte ve profil hesabına katılmaktadır. Yapı tahmininde en yaygın kullanılan profil matrisi türetme yöntemi PSI-BLAST'dır. Bunun çeşitli sebepleri arasında programın hızlı çalışması, belirli seviyede hassaslık (İng. sensitivity) sağlaması, yazılımın kolay erişilir olması ve düzenli olarak güncellemesi sayılabilir. Ancak PSI-BLAST yöntemi daha uzak protein benzerliklerini bulabilse de hatalı eşleşmeler de yapabildiğinden bu yöntem ile türetilen profil matrisleri belirli bir gürültü içermektedir. Bu gürültü, yapısal özelliklerin tahmin edilmesini zorlaştırıcı bir rol oynasa da yöntem yaygın olarak kullanıldığından ve belirli bir seviyede doğruluk oranına sahip olduğundan bu projede kullanılacak ilk profil matrisi PSI-BLAST yöntemi ile elde edilmiştir. PSI-BLAST ile profil matrisleri elde edildikten sonra sigmoid dönüşümü ile 0 ile 1 arasında haritalanmıştır.

3.1.2 HHMAKE PSSM

Dizi hizalama algoritmalarıyla bulunan proteinler ile çoklu hizalama yapıldıktan sonra saklı Markov modellerine (HMM) dayanan profiller de türetilebilir ve yinelemeli olarak profil dizi hizalaması ya da profil profil hizalaması için kullanılabilir (Remmert vd., 2012). Saklı Markov modellerine dayanan profillerin standart profillere göre daha hassas olduğu ve daha uzak protein benzerliklerini keşfedebildiği bilinmektedir. Bu projede HHBlits yöntemiyle ilk yinelemeden elde edilen saklı Markov modellerinin eşleşme düğümlerindeki dağılımlar (match state distribution) doğrusal ölçekleme (linear scaling) ile pozisyona özel skor matrislerine (PSSM) dönüştürülmüş ve ikinci tip profil matrisi olarak kullanılmıştır.

3.1.3 Yapısal Profil Matrisleri

İkincil yapı tahmini için amino asit dizilerinin çoklu hizalanmasına dayanan profil matrislerine ek olarak yapısal profil matrisleri de öznitelik olarak kullanılmıştır (Li vd., 2011; Zhou vd., 2009). Burada dizi hizalama yöntemleriyle bulunan proteinlerin yapısal dizileri (örn. ikincil yapı dizisi) kullanılarak yapısal profil matrisleri oluşturulabilmektedir. Tipik olarak ikincil yapı tahmini için oluşturulan bir yapısal profil matrisinin boyutu 3×U'dir (U hedef proteindeki amino asit sayısıdır ve her sütunda o amino asit için üç ikincil yapı durumundan birisinin gözlemlenme skoru bulunmaktadır). Yapısal profillerde hedef proteine benzerliği olan kalıp proteinlerin yapı bilgisini de kullandığından sadece dizi profil kullanan yöntemlerden ayrı kategoride değerlendirilebilir. Bir diğer kategori de hedef protein dizisel olarak benzerliği belirli bir seviyenin üzerinde olan (örn. %50'den daha fazla) kalıp proteinlerin ikincil yapı bilgisinin doğrudan doğruya tahmin için kullanılmasıdır. Bu durumda yapısal profillerin

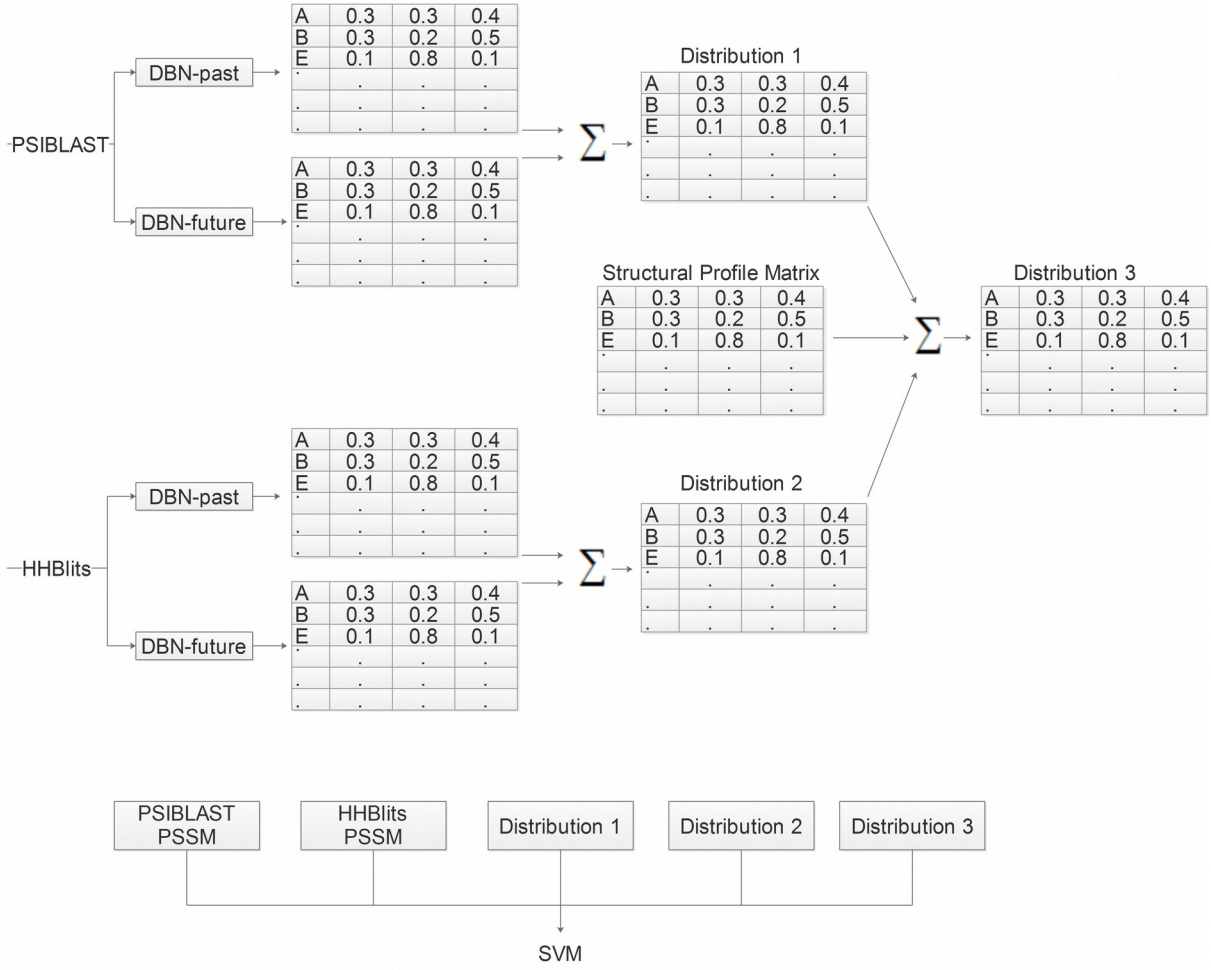
kullanıldığı durum, dizisel profillerin kullanıldığı durum ile doğrudan doğruya kalıp proteinlerin tahmin için kullanıldığı durumun arasında bir kategori olarak düşünülebilir.

Yapısal profillerin oluşumunda kullanılan protein dizileri, hedef proteine ne kadar çok benzerse tahminlerin başarısı da o kadar iyileşmektedir. Ayrıca hedef proteinin tamamı yerine bir alt bölgesine benzediği (yerel benzerlik) durumlarda da bir boyutlu yapı tahmininde bir miktar iyileşme olabilmekte, benzeyen bölge daha uzun oldukça tahminlerde daha belirgin bir iyileşme sağlanabilmektedir. Bu projede yapısal profil matrislerinin DSPRED yöntemine dahil edilmesi ile tahmin başarı oranında önemli oranda iyileşme elde edilmiştir.

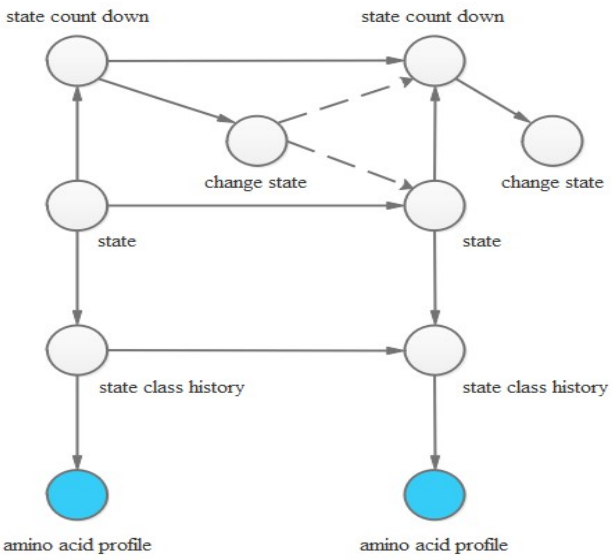
3.2 DSPRED Tahmin Yöntemi

DSPRED yöntemi, ikincil yapı, dihedral açılar ve çözücü erişilirlik gibi bir boyutlu yapısal özelliklerin tahmini için proje yürütücüsü tarafından geliştirilmiş iki aşamalı bir hibrit sınıflandırıcıdır. DSPRED yönteminin basamakları Şekil 1'de gösterilmektedir. DSPRED'de PSI-BLAST (Altschul vd., 1997) ve HHBlits (Remmert vd., 2012) yöntemlerinden elde edilen konuma özgü puanlama (profil) matrisleri için ayrı dinamik Bayes ağları (DBN'ler) eğitilmiştir. Bu girdi öznitelikleri sırasıyla PSIBLAST PSSM ve HHMAKE PSSM olarak ifade edilmiştir.

DSPRED yönteminde iki tür dinamik Bayes ağı modeli mevcuttur. Burada DBN-geçmiş verilen bir pozisyondaki profil vektörünün kendinden önceki pozisyonlara bağlı olduğu modeli, DBN-gelecek ise verilen bir pozisyondaki profil vektörünün kendinden sonra gelen pozisyonlara bağlı olduğu modeli temsil eder. Profil vektörleri profil matrisinin sütunlarıdır ve amino asitlerin sayısı kadar sütun bulunur. Her DBN'nin çıktısı, girdi özellikleri verilen ikincil yapı sınıfı etiketleri için bir marjinal posteriori olasılık dağılımıdır. Daha sonra bu dağılımlar çeşitli kombinasyonlarda ortalama alma işlemi ile birleştirilir. Örneğin, Dağılım 1, PSI-BLAST PSSM'leri kullanan DBN'ler tarafından üretilen dağılımların ortalamasını temsil eder; Dağılım 2, HHMAKE PSSM özelliklerini kullanan DBN'ler tarafından üretilen dağılımların ortalamasını temsil eder ve Dağılım 3, Dağılım 1'in, Dağılım 2'nin ve HHBlits yöntemi kullanılarak elde edilen yapısal profil matrislerinin ortalaması alınarak hesaplanır. Bu projede, ikincil yapı tahminlerinde ikincil yapı sınıflarının sayısı üç olduğundan, Dağılım 1, 2 ve 3 boyutları 3xU'dur, burada U amino asitlerin sayısıdır. Sonuç olarak, her sütun, bir amino asit pozisyonundaki ikincil yapı sınıflarının tahmini ihtimallerini içerir. DSPRED'in ikinci aşamasında, profil matrisleri (PSI-BLAST ve HHMAKE), Dağılım 1, 2 ve 3 ile birleştirilir ve destek vektör makinesi gibi ayırt edici bir sınıflandırıcıya gönderilir. Bunun için yapı sınıfının öngörüleceği her amino asit etrafında simetrik bir pencere alınır ve bu sütunlardaki öznitelikler girdi vektörünü oluşturmak üzere birleştirilir. İkinci aşamadaki sınıflandırıcı, pencerenin ortasındaki amino asidin yapı etiketini tahmin etmek için kullanılır ve ilk aşamadaki DBN sınıflandırıcıların hatalarını düzeltmeyi amaçlar.



Şekil 1. Bir boyutlu protein yapı tahmini için DSPRED yönteminin adımları



(A)

State	H	H	H	H	H	L	L	L	H	H	H	E	E	E	E	E	L	L	L	L	L	L
state count down	5	4	3	2	1	3	2	1	3	2	1	5	4	3	2	1	5	5	4	3	2	1
change state	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0

(B)

Şekil 2. (A) Bir boyutlu yapı tahmini için dinamik Bayes ağı. (B) İkincil yapı parçalarını modellemek için kullanılan değişkenler.

Üç durumlu ikincil yapı sınıfını tahmin etmek için kullanılan dinamik Bayes ağ modeli Şekil 2 (A)'da gösterilmiştir (Aydin vd., 2011). Dinamik Bayes ağı (DBN) üretken (generative) bir modeldir ve saklı Markov modelinin (HMM) üst kümesidir (Reynolds vd., 2008). DBN'ler, belirli olasılık kurallarına uyarak gizli sınıf değişkenlerinden profil vektörlerinin üretilmesini modellemektedir. Şekil 2 (A)'daki DBN düğümleri rastgele değişkenleri temsil etmektedir. Durum (state) değişkeni, bir amino asidin ikincil yapı sınıfı etiketini temsil eder. Amino asit profili (amino acid profile) değişkeni, bir amino asidin profil vektörünü içerir ve eğitim ve test sırasında gözlemlenir. Bu vektörler PSI-BLAST ve HHMAKE tarafından üretilen PSSM'lerin sütunlarına karşılık gelir. Bunlara ek olarak geçerli ve önceki ikincil yapı etiketleri birleştirilir ve durum sınıfı geçmişinin olası her değerine farklı bir koşullu Gauss dağılımı sığdırmak için kullanılan durum sınıfı geçmişi (state class history) değişkeninde saklanır. Bu koşullu dağılım, durum sınıf geçmişi değişkeni göz önüne alındığında amino asit profilini gözlemleme olasılığıdır ve girdi özelliklerinin üretilmesinden sorumludur. Durum geri sayımı (state count down), geçerli konumdan bir sonraki ikincil yapı parçasına kadar bir mesafe değeri içerir. Bu değişken, bölümlerin uzunluk dağılımını modellemeye yardımcı olur. Geçerli konumdan bir sonraki bölüme kadar (N_A ile gösterilen) amino asitlerin sayısı bir eşikten (D_{max} olarak adlandırılır) azsa, durum sayımı N_A 'ya eşit olur, aksi halde D_{max} olarak ayarlanır. Durum geri sayımının D_{max}'dan daha az olduğu pozisyonlar için, uzunluk dağılımı, olasılık sıklıklarını kullanan maksimum olasılık (maximum likelihood) yaklaşımı kullanılarak tahmin edilir. Kalan pozisyonlar için, uzunluk dağılımı bir geometrik dağılım ile modellenir. Durum değiştirme değişkeni ise (change state) bir segmentten diğerine geçiş için kullanılır. DBN modelleri, bilinen yapı etiketlerine sahip proteinleri kullanarak eğitildikten sonra, sınıf etiketlerinin marjinal ve sonraki olasılıklarını en üst düzeye çıkaran tahminler verimli algoritmalarla hesaplanabilir. Bu projede DBN modellerini oluşturmak için Linux tabanlı GMTK yazılım paketi (Graphical Models Toolkit) kullanılmıştır. GMTK, model eğitimi için EM algoritmasını ve tahminler için junction tree algoritmasını kullanır.

3.3 İkinci Aşama İçin Sınıflandırma Modelleri

Bu projede yürütücü tarafından daha önce geliştirilen DSPRED hibrit sınıflandırıcısının ikinci aşaması için destek vektör makinası, deep CNF ve rastgele orman yöntemleri eğitilmiş ve elde edilen tahminler bir topluluk yaklaşımı ile birleştirilmiştir. Deep CNF yöntemi derin

konvolüsyonel ağlar (deep convolutional networks) ile koşullu rastgele alan (conditional random field) yöntemini birleştiren hibrit bir yöntemdir (Wang vd., 2016).

3.3.1 Destek Vektör Makineleri

Destek vektör makinesi (DVM), makine öğrenmede ayırt edici (discriminative) sınıflandırıcılar arasındadır. DVM'nin temel amacı, iki veya daha fazla sınıf arasında en uygun ayırımı yapan bir hiper düzlemi tanımlamaktır.

DVM hem doğrusal hem de doğrusal olmayan veri kümelerini sınıflandırabilir. İki sınıfın olduğunu varsayalım. Bu sınıfların veri örneklerini birbirinden ayıran sonsuz sayıda düzlem çizilebilir. Bu noktada, DVM'nin amacı, düzleme en yakın örnekler arasındaki mesafeyi en yükseğe çıkaran hiper düzlemi bulmaktır. Bu projede destek vektör makinaları libSVM programı ile gerçeklenmiştir.

3.3.2 Rastgele Orman

Rastgele orman, sınıflandırma veya regresyon için kullanılan karar ağaçlarının topluluğunu oluşturan bir yöntemdir. Biyomedikal, fizik, sağlık ve biyoinformatik gibi birçok farklı probleme uygulanır. Bagging tekniğini uygulayan modeller arasındadır ve ağırlıklı oy çoğunluğu ile temel öğrenicilerinin kararlarını birleştirir. Her ağaç, orijinal öznitelik kümesinden rastgele seçilen farklı bir alt kümesiyle eğitilir. Ayrıca her ağaç modeli, önyükleme örneklemesi ile elde edilen biraz daha farklı bir eğitim seti kullanılarak oluşturulmuştur. Karar ağaçları oluşturmak için, Gini dizini katkı ölçüsü olarak kullanılır. Bu projede rastgele orman yöntemi WEKA programı ile gerçeklenimiştir.

Rastgele ormanlar çeşitli nedenlerle tercih edilir. Overfitting’e karşı dayanıklıdırlar. Öznitelik sayısının artırılması doğrudan doğruya overfitting’e neden olmaz. Veri kümesinin boyutu arttıkça, eğitim setindeki tüm kavramları öğrenmek için ağaçların sayısı artabilir. Bu nedenle, ağaçların sayısı ve rastgele seçilen özelliklerin sayısı doğru seçilirse, rasgele bir orman, overfitting probleminden kaçınabilir.

Rasgele ormanların diğer avantajları arasında, büyük verilerin etkin performansı, parametrelerin ayarlanması, eksik değerlerin ele alınması gibi kolaylıklar bulunmaktadır.

3.3.3 Derin Katlamalı Sinir Alanları

Derin öğrenme, verilerdeki doğrusal olmayan ve karmaşık ilişkileri öğrenmeyi amaçlayan bir makine öğrenme tekniğidir. Görüntü tanıma, biyoinformatik, doğal dil işleme gibi birçok alana uygulanmıştır. Derin öğrenme yaklaşımlarının amacı, daha basit ve daha düşük seviyedeki özellikleri kullanarak daha üst düzey ve daha karmaşık nitelikler öğrenmektir.

Koşullu sinirsel alanlar, koşullu rastgele alanların girdi özellikleri ile çıktı katmanı arasındaki doğrusal olmayan bağlantıyı çözme probleminden ötürü geliştirilmiştir. Derin katlamalı sinir alanı (CNF), derin katlamalı sinir ağları ile koşullu rastgele alan (CRF) birleştirilerek elde edilen hibrit bir sınıflandırıcıdır. Derin bir öğrenme tekniği olarak, giriş özellikleri ile çıktı arasında doğrusal olmayan bağlantıları elde etmeyi sağlar. Ayrıca koşullu rastgele alan modelinin son katmanındaki komşu çıktı etiketleri arasındaki bağıntıyı modelleyebilir. Sonuç olarak, katlamalı sinir ağları ile koşullu rastgele alanların avantajlarını tek bir modelde bir araya getirir (Wang vd., 2016).

3.4 Çapraz Doğrulama ile Model Değerlendirmesi

Makine öğrenme ve veri madenciliğinde kullanılan veri kümeleri genellikle tahmini modelleri değerlendirmek için eğitim ve test kümesi olarak iki kısma ayrılmıştır. Eğitim seti, modeli eğitmek ve test seti doğruluğunu değerlendirmek için kullanılır. K-katlı çapraz doğrulama, veri kümesini k eşit boyutlu alt kümelere bölerek tahmini modeli test etmek için kullanılan bir tekniktir. Her yinelemede, k alt kümelerinden biri test kümesi olarak kullanılırken diğer k-1 alt kümeleri eğitim kümesini oluşturur. Bu işlem, tüm alt kümeler test seti olarak kullanılana kadar k defa tekrarlanır. Bu projede, yöntemlerin doğruluğunu değerlendirmek için farklı sayılarda çapraz doğrulama kullanılmıştır. Çapraz geçerlilik doğruluğunu değerlendirmeden önce, modellerin hiper parametreleri, çapraz geçerlilik tekrarlaması için ayrı ayrı optimize edilir. Bu amaçla, her bir eğitim setindeki proteinlerin % 10'u rasgele olarak doğrulama (validation) kümesi olarak seçilir. Eğitim setinin geri kalan %90'ı ve doğrulama kümesi, modellerin hiper parametrelerini optimize etmek için kullanılır. Optimizasyon işlemi tamamlandıktan sonra, eğitim kümesindeki tüm proteinler modelleri öğrenmek için kullanılır ve tahminler karşılık gelen test setinde hesaplanır. Bu işlem her kat için tekrarlanır ve her model için doğruluk oranı elde edilir. Genel doğruluk oranı ise, bu elde edilen doğruluk değerlerinin ortalaması alınarak hesaplanır.

3.5 Parametre Optimizasyonu

Bir sınıflandırıcının performansını etkileyebilecek birden fazla koşul vardır. Bunlardan biri düzenli eğitim sürecinde doğrudan öğrenilemeyen doğru hiper parametrelerin seçimidir. Bu parametrelerin en iyi duruma getirilmesi, modelin karmaşıklığını düzenlememizi ve fazla uyumun yanı sıra az oranda uyumu önlememize de olanak tanır.

Optimizasyon için sınıflandırıcı modelleri, bir eğitim grubundaki proteinlerin %90'ını içeren ve rastgele seçilmiş bir alt kümedeki çeşitli hiper parametre kombinasyonları için eğitilir ve ilgili doğrulama kümesindeki tahminler hesaplanır. Doğrulama kümesindeki genel doğruluk oranını maksimuma çıkaran hiper parametreler optimum değerler olarak seçilir.

3.5.1 Destek Vektör Makineleri için Parametre Optimizasyonu

DVM için, C ve Gammayı (γ) optimize ettik. C, her destek vektörünün etkisini kontrol eden bir maliyet fonksiyon parametresidir. C için uygun bir değer seçmek, hatayı tolere etmek için önemlidir. C'nin değeri düşük olduğunda, karar daha düzgün olur ve marjin artar. C değeri yüksek olduğunda, marjindeki azalma nedeniyle öğrenme aşamasındaki duyarlılık artar.

Gamma parametresi yayılma etkisinin miktarını belirler. Gamma parametresinin değeri küçük olduğunda, karar sınırı doğrusal hale gelir. Gamma daha yüksek değer aldığında, karar sınırı doğrusal olmayan bir hal alır.

SVM optimizasyonu için bir C ve gamma değerleri şeması düşünülüp her parametre için aşağıdaki değerler seçilmiştir.

C = (2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}, 2^{7}, 2^{9}, 2^{11}, 2^{13})

Gamma = (2^{-15}, 2^{-13}, 2^{-11}, 2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3})

Sonuç olarak, dikkate alınması gereken toplam 100 farklı kombinasyon oluşmaktadır. Belirli çapraz doğrulama yinelemesi için (onaylama setinde en iyi doğruluğa sahip olan) optimum kombinasyon bulunursa, model tam eğitim setinde eğitilir ve tahminler test setinde hesaplanır. Bu prosedür, çapraz doğrulama denemesinin diğer yinelemeleri için tekrarlanır.

3.5.2 Rasgele Orman için Parametre Optimizasyonu

Rasgele orman için en iyi duruma getirilmiş tek parametre, ağaçların sayısıdır (WEKA'da yineleme sayısı olarak da adlandırılır). Aşağıdaki değerler, çapraz doğrulamadaki her bir kat için dikkate alınır.

Ağaç sayısı = (5 10 15 20 25 50 75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500).

3.5.3 Derin Katlamalı Sinir Alanları için Parametre Optimizasyonu

Derin katlamalı sinir alanları için optimize edilen parametreler, gizli katmanların sayısı, her bir katmandaki gizli birim sayısı, her bir gizli katmanda uygulanan iki boyutlu çekirdek penceresinin genişliği ve normalleştirme katsayısıdır. Pencere boyutu, pencere dizisi adı verilen bir değişkenle belirtilir. Örneğin, pencere dizisi 5 olduğunda, çekirdek penceresinin boyutu her katmanda 11 olacaktır. Gizli düğümlerin sayısı ve gizli katmanların sayısı, düğüm dizisi adı verilen bir değişkenle belirtilir. Örneğin, düğüm dizesi "50,50,50,50,50" ise, ağ her katmanda 50 gizli düğüm bulunan beş gizli katmanı içeriyor. Derin KSA'de L_2-normu normalleştirme için kullanılır. Aşağıdaki değerler, optimizasyon için kullanılan parametre değerleridir.

Gizli Katman Sayısı = (3, 4, 5)

Gizli Birim Sayısı = (75, 100, 125)

Çekirdek Pencere Boyutu = (3, 4, 5)
Düzenleme Katsayısı = (10, 50, 100)

3.6 Topluluk Yöntemi

Topluluk yöntemlerinin ardındaki temel fikir, tahmin doğruluğunu ve modellerin sağlamlığını arttırmak amacıyla birkaç sınıflandırıcının tahminlerini birleştirmektir. Bireysel sınıflandırıcıların çıktıları birleştirilerek hesaplanan bir yöntem oluşturulup bu yöntemin nihai sınıflandırıcı performansını iyileştirilebileceği bulunmuştur.

3.6.1 Model Ortalaması

Model ortalamaları, bir ortak tahmin oluşturmak için birden fazla yöntemden elde edilen tahmin skorlarının ağırlıklı toplamını hesaplar. Sınıflandırıcıları model ortalamaları ile birleştirmek, genel olarak, varyansı azaltma olasılığı nedeniyle, herhangi bir sınıflandırıcıdan daha iyi sonuçlar elde eder. Model ortalaması şu şekilde formüle edilmiştir:

$$p^c(y|X)=\frac{1}{N}\sum_{n=1}^{N} p^n(y|X) \qquad (15)$$

Burada X bir amino asidin girdi özellik vektörünü, y çıktı sınıf etiketini (helezon, iplikçik veya döngüyü temsil eden H, E veya L çıktı sınıf etiketi), $p^c(y|X)$verilen sınıf vektörünün özellik vektörünün posteriori olasılığını, n toplulukta kullanılan modelin endeksini, $p^n(y|X)$, n sayıdaki modelinden girdi vektörü verilen sınıf etiketinin posteriori olasılığını, N ise topluluktaki modellerin sayısını temsil etmektedir. Son sınıf tahmini, $p^c(y|X)$ i maksimum yapan ikincil yapı etiketlerini seçerek elde edilir. Bu projede, DVM, derin katlamalı sinir alanları ve rasgele orman tarafından elde edilen tahminleri birleştirmek için model ortalamaları uygulanmıştır. Aşağıdaki kombinasyonlar birleştirme amaçlı kullanılmıştır.

- DVM + rasgele orman
- DVM + derin katlamalı sinir alanları
- Derin katlamalı sinir alanları + rasgele orman
- DVM + derin katlamalı sinir alanları + rasgele orman

3.7 Boyut Düşürme Yöntemleri

Protein yapı tahmininde yeni öznitelik çıkarımı yaklaşımları geliştirilmektedir ancak bunlar birbirini tamamlayıcı nitelikte olduğundan genellikle problem için tek başına yeterli olmamaktadır. Bu özniteliklerin birlikte kullanılması durumunda ise veri uzayının boyutu artmaktadır. Bir veri kümesindeki öznitelik (boyut) sayısı gerekenden fazla ise model eğitme zamanı artar ve eğitilen modeller aşırı uyum davranışına (overfitting) takılabilir. Bunun sonucunda yöntemin tahmin başarı oranı düşebilir. Ayrıca bazı öznitelikler gürültü içerebildiğinden arzulanan örüntü ve bilgiler yeterince iyi öğrenilemeyebilir. Diğer taraftan

gerekenden az öznitelik bulunması bu bilgilerin öğrenilmesi için yetersiz kalabilir. Dolayısıyla bir makine öğrenmesi yönteminde doğru sayıda öznitelik bulunması sınıflandırma başarısı için önem taşımaktadır. Eğer oluşturulan öznitelik kümesinin boyutu gerekenden fazla ise izlenebilecek yaklaşımlardan birisi boyut düşürmedir. Bunun için de iki temel yaklaşım izlenebilir: izdüşüm tabanlı haritalama (projection based mapping) ve öznitelik seçimi (feature selection). Bu projede her iki kategoride bulunan yöntemler gerçeklenmiş, elde edilen öznitelik kümeleri kullanılarak eğitilen modellerin tahmin başarı oranları karşılaştırılmıştır. İlk olarak öznitelikleri sıralamak için kullanılan çeşitli skorlama teknikleri anlatılacak daha sonra en uygun öznitelik kombinasyonunu bulmak için kullanılan arama algoritmaları açıklanacaktır.

3.7.1 Ki-Kare Skoru

Randy Kerber tarafından 1992 yılında ortaya atılan ve Huan lui ve Rudy Setiono tarafından 1995 yılında geliştirilen χ^2 testi olarak da bilinen Ki-kare yöntemi değişkenlerin veri setini tanımlamaya uygun olup olmadığını belirlemek içinde kullanılabilir (Kavzoğlu vd., 2014). Ki-kare testinde H_0 ve H_1 olmak üzere iki hipotez bulunmaktadır. H_0 veri setindeki değişkenlerin uygun olduğu, H_1 ise veri setindeki değişkenlerin uygun olmadığı hipotezidir. İki aşaması olan bu testin ilk aşamasında gözlenen değerlerin gerçek sınıflara göre ki-kare (χ^2) istatistiği hesaplanır. χ^2 değeri sıfır ile pozitif sonsuz arasında değerler alabilir. Bu değerin sıfıra yaklaşması gözlenen frekans değerleri ile beklenen frekans değerlerinin daha uyumlu olduğunu gösterir. Bu değerin çok büyük olması uyumsuzluğu işaret etmektedir. Bu nedenle testin ikinci aşamasında ilk aşamada hesaplanan χ^2 değeri Ki-kare dağılımındaki belirlenen eşik değeri ile kıyaslanır. Bu eşik değeri önemlilik seviyesine ve serbestlik derecesine bakılarak bulunmaktadır. Önemlilik seviyesi testi yapan kişi tarafından belirlenen yüzde değeridir, serbestlik derecesi ise veri setindeki öznitelik sayısının bir eksiğidir. Hesaplanan değer belirlenen değerden büyük ise H_1 hipotezi, küçük ise H_0 hipotezi kabul edilir. Eşitlik 1'de Ki-kare istatistiğinin nasıl hesaplandığı formülize edilmiştir.

$$\chi^2 = \sum_{i=1}^{n} \frac{(o¿¿i - e_i)^2}{e_i} ¿ \quad (1)$$

Bu eşitlikte n veri setindeki öznitelik sayısını, o_i i'ninci öznitelik için gözlenen frekans değerini, e_i ise i'ninci öznitelik için beklenen frekans değerini temsil etmektedir.

3.7.2 Bilgi Kazancı Skoru

Bilgi kazancı, veri seti özniteliklere göre bölündüğünde tahmini kaybı hesaplamada kullanılan entropiye dayalı yöntemlerden biridir. Entropi sistemin düzensizliğini ya da belirsizliğini

belirleyen 0 ile 1 arasında bir değerdir. Entropi değerinin 1'e yaklaşması sistemin daha çok bilgi içerdiğini göstermektedir. Bilgi kazancı yönteminin ilk aşamasında, verilen bir veri setinin sınıf etiketleri için entropi değeri eşitlik 2'de formülize edildiği gibi hesaplanır.

$$E = -\sum_{i=1}^{n} ¿¿¿ \quad (2)$$

Bu eşitlikte n sınıf sayısını, $ns(i)$ i'ninci sınıf için örnek sayısını, N ise toplam örnek sayısını temsil etmektedir.

Bu yöntemin ikinci aşamasında veri setindeki her bir öznitelik için entropi değeri hesaplanır ve bu yeni entropi değeri ilk aşamada bulunmuş olan değerden çıkarılarak bilgi kazancı hesaplanır. Bilgi kazancı veri setinin bölünme sonrası temsil değerini göstermektedir. Bu nedenle bu değerin büyük olması beklenmektedir. Bilgi kazancı yöntemi ile öznitelik seçimi yapılırken sistemi tanımlamada yetersiz kalan değişkenler veri setinden çıkarılır ve kalan değişkenler sistemi eğitmek için kullanılır. Eşitlik 3'te her bir öznitelik için entropi değerinin, eşitlik 4'te ise bilgi kazancı değerinin hesaplanması formülize edilmiştir.

$$E(i) = \sum_{k=1}^{n} \frac{ns(k)}{N} * \sum_{m=1}^{nc} \frac{-nsc(k,m)}{ns(k)} \left(\log_2 \frac{nsc(i,m)}{ns(i)}\right) \quad (3)$$

$$B(i) = E(i) - E \quad (4)$$

Eşitliklerde $E(i)$ i'ninci öznitelik için entropi değerini, n i'ninci özniteliğin alabileceği farklı değer sayısını, $ns(k)$ i'ninci özniteliğin k değerine ait örnek sayısını, N veri setindeki toplam örnek sayısını, nc veri setindeki sınıf sayısını, $nsc(k,m)$ i'ninci özniteliğin k değerine ait m sınıfını temsil eden örnek sayısını, $B(i)$ bilgi kazancını, E ise eşitlik 2'de hesaplanan entropi değerini temsil etmektedir.

3.7.3 Kazanım Oranı Skoru

Kazanım oranı, bilgi kazancı yöntemine alternatif olarak aynı amaç doğrultusunda kullanılan öznitelik seçim yöntemlerinden biridir. Veri setinde bir öznitelik çok fazla farklı değere sahip olduğunda o öznitelik için her bir farklı değere düşen örnek sayısı düşük olmaktadır. Bu nedenle o öznitelik için hesaplanan entropi değeri küçük, bilgi kazancı ise büyük çıkmaktadır. Bilgi kazancı yönteminde de anlatıldığı gibi bu değerin büyük çıkması o değişkenin veri setini tanımlamada iyi olduğunu göstermektedir. Özniteliğin alabileceği farklı değer sayısının çok olması durumunda bilgi kazancı yönteminin o özniteliği iyi bir seçici olarak seçmesi sistemin ezberleme yapmasına neden olabilmektedir. Kazanım oranı bu probleme çözüm olarak her bir öznitelik için bölünme bilgisini hesaplar ve elde edilen bilgi kazancını bölünme bilgisine bölerek kazanım oranını hesaplar. Bu oranın 1'e yaklaşması o değişkenin veri setini tanımlamada başarılı olduğunu göstermektedir. Eşitlik 5'te bölünme bilgisi, eşitlik 6'da ise kazanım oranı verilmiştir.

$$S(i) = -\sum_{k=1}^{n} \frac{ns(k)}{N} * ¿(\log_2 \frac{ns(k)}{N}) \quad (5)$$

$$K(i) = \frac{B(i)}{S(i)} \quad (6)$$

Bu eşitliklerde $S(i)$ i'ninci öznitelik için bölünme bilgisini, n i'ninci özniteliğe ait farklı değer sayısını, $ns(k)$ i'ninci özniteliğin k değerine ait örnek sayısını, N toplam örnek sayısını, $K(i)$ i 'ninci öznitelik için kazanım oranını $B(i)$ ise eşitlik 4'te hesaplanan bilgi kazancı değerini temsil etmektedir.

3.7.4 Minimum Fazlalık Maksimum İlgi

Minimum fazlalık maksimum ilgi yaklaşımında gereksiz öznitelikler elenerek veri setini en iyi tanımlayan öznitelikleri seçmek amaçlanmaktadır. Bunu yaparken birbiriyle ilgili özniteliklerin de birlikte seçilmesi arzulanır. Bu yöntemin ilk aşamasında karşılıklı bilgi değeri eşitlik 7 kullanılarak hesaplanmaktadır.

$$I(X,Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} p(x_i, y_j) \times \log \frac{p(x_i, y_j)}{p(x_i) \times p(y_j)} \quad (7)$$

Bu eşitlikte n değeri veri setindeki örnek sayısını, $p(x_i, y_j)$ bağımlı olasılık dağılım değerini, $p(x_i), p(y_j)$ ise marjinal olasılık değerini temsil etmektedir. Daha sonra karşılıklı bilgi değeri kullanılarak mRed ve mRel değerleri aşağıdaki eşitliklerde formülize edildiği gibi hesaplanmaktadır.

$$mRed = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} I(x_i, y_j) \quad (8)$$

$$mRel = \frac{1}{m} \sum_{i=1}^{m} I(x_i, h_i) \quad (9)$$

Bu eşitliklerde m yeni veri setinin boyunu h_i ise i'inci örnek için sınıf etiketini temsil etmektedir. Bu yöntemin son aşamasında $mRel - mRed$ ve $\left(\frac{mRel}{mRed}\right)$ değerlerini maksimum yapan öznitelik seti seçilerek yöntem sonlandırılmaktadır.

3.7.5 Genetik Arama Algoritması

Genetik algoritma (GA) doğal seleksiyon, çaprazlama ve mutasyon tabanlı, biyolojiden ilham alan global arama optimizasyon tekniğidir. Bu yöntemde öncelikle aday çözümlerden oluşan bir popülasyon üretilir ve bu popülasyon belirlenen durdurma kriteri sağlanıncaya kadar

seleksiyon, çaprazlama ve mutasyon adı verilen genetik işlemler aracılığı ile güncellenir. GA daha iyi çözümleri bulma sürecinde en iyi olanın hayatta kalması fikrini kullanır. GA tek bir çözümü kademeli olarak değiştirmektense bir çözüm popülasyonunu güncelleyerek arama yapması yönüyle geleneksel doğrusal olmayan optimizasyon tekniklerinden ayrılır. Klasik optimizasyon algoritmaları iterasyon noktalarının yerel özellikleri ile ilgilendiği için kolayca yerel ekstremum noktalarına takılabilirler. Bunun aksine GA sistematik aramaya ek olarak rasgele arama operatörü de kullandığından dolayı yerel minimum veya maksimum noktasına takılması önlenmiş olur.

GA optimize edilecek parametrelerin bir dizi çözümüyle başlar. Çözümü oluşturan kromozomların her bir parametresi gen olarak adlandırılır ve parametreler ikili bit dizisi, tam ya da reel sayı şeklinde kodlanabilirler. Herhangi bir ön bilgi olmadığında ilk popülasyondaki her kromozom, düzgün dağılım kullanarak rastgele oluşturulur. Daha sonra her bir çözümün uygunluğunu belirlenen bir fonksiyon yardımı ile bulunurak büyükten küçüğe doğru bir sıralama yapılır. Bu sıralanan nesillerin yardımı ile mutasyon, çaprazlama gibi teknikler kullanılarak yeni nesiller üretilir ve istenilen başarı oranı elde edilene kadar bu işlemler yeni nesiller üzerinde de tekrar edilir. Bu özelliği ile genetik algoritma ile, bir çok algoritmaya göre daha yavaş çalışmasına rağmen daha yüksek başarı oranı elde edilir. Burada genetik algoritma optimum öznitelik kombinasyonunu bulmak için kullanılan bir arama tekniğidir.

3.7.6 Açgözlü Arama Algoritması

En iyi sonucu veren öznitelikleri bulmayı hedefleyen bu yöntemin ileriye doğru ve geriye doğru olmak üzere iki farklı versiyonu bulunmaktadır. İleriye doğru öznitelik seçiminde (FFS) öznitelik seçme işlemi azdan çoğa doğru yapıldığı için ileriye doğru olarak adlandırılmıştır. Yöntemde ilk olarak boş bir öznitelik seti ile başlanır. Daha sonra hedef hipotezi H_0 ve bunun karşıt hipotezi H_1 belirlenir. H_0; başarı oranını artırmak, doğru tespit oranı artırmak gibi isteğe bağlı olarak herhangi bir performans unsuru olabilmektedir. Ardından D özniteliği olan bir veri setinden bir öznitelik seçilerek oluşturulan sete eklenir ve eğitim yapılır. Eğer H_0 hipotezi sağlanıyorsa eğitim setine seçilmiş olan öznitelik eklenerek, H_1 hipotezi sağlanıyorsa oluşturulan eğitim seti değiştirilmeden aynı işlemler diğer öznitelikler içinde yapılır. D adet özniteliğin hepsi için aynı işlem tekrarlanarak yöntem sonlandırılır ve en uygun sonucu veren veri seti tespit edilmeye çalışılır.

Geriye doğru öznitelik seçimi (BFS); FFS ile aynı amaç doğrultusunda kullanılan bir yöntemdir. FFS'den farklı olarak öznitelik eleme işlemi geriye doğru yani çoktan aza olacak şekilde yapılmaktadır. Bu nedenle yöntemde ilk olarak D adet özniteliği olan veri setinin tümü öznitelik seti olarak belirlenir. Daha sonra FFS'ye benzer şekilde H_0 ve H_1 hipotezleri

belirlenir. Üçüncü adımda ise belirlenen öznitelik setinden bir öznitelik çıkarılarak eğitim işlemi yapılır. H_0 hipotezi sağlanırsa o öznitelik oluşturulan setten çıkarılarak, H_1 sağlanıyorsa öznitelik setinde değiştirilmeden işlemler tüm öznitelikler için tekrar edilir ve model sonlandırılır. FFS ve BFS yöntemlerinin ikisi de en uygun öznitelik setini bulmayı hedefler ancak bulunan öznitelik setinin en uygun set olduğunu garanti edemezler.

3.7.7 En İyi İlk Önce Arama Algoritması

En iyi ilk önce algoritması açgözlü algoritmaya benzer bir şekilde çalışmaktadır. Bu algoritmanın açgözlü algoritmadan tek farkı algoritmanın istenilen noktadan başlayarak istenilen doğrultuda (ileriye ya da geriye doğru) gidebiliyor olmasıdır.

3.7.8 Temel Bileşen Analizi

Temel bileşen analizi (TBA) (İng. principal component analysis), değişkenler arasındaki bağımlılığın bulunması için kullanılan bir boyut düşürme tekniğidir. En büyük varyans ile en az kaybı hedefleyen bu yöntemin ilk aşamasında her bir değişken için diğer değişkenlerle olan kovaryans değeri eşitlik 10'da formülize edildiği gibi hesaplanır. Kovaryans değeri iki değişkenin birlikte değişimini temsil eden bir değerdir. Bu değerin pozitif olması iki değişkenin aynı anda büyüdüğü ya da küçüldüğü, negatif olması iki değişkenden biri büyür iken diğerinin küçüldüğü, sıfır olması ise bu iki değişkenin birbirinden bağımsız olduğu durumu belirtir.

$$cov(X,Y)=\frac{\sum_{i=1}^{n}\left(X_i-\acute{X}\right)\times\left(Y_i-\acute{Y}\right)}{n-1}$$

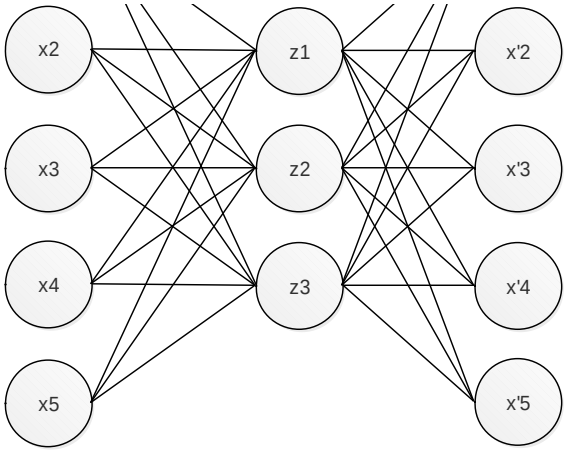
(10)

Bu eşitlikte X,Y her bir değişkeni; n örnek sayısını; X_i,Y_i ilgili değişken için i'ninci örnek değerini; $\overline{X}$, $\overline{Y}$ ilgili değişken için verilen örneklerin ortalama değerini temsil etmektedir. Bu adımın devamında bulunan kovaryans değerleri kullanılarak kovaryans matrisi oluşturulur. Sonraki adımda ise bu matris kullanılarak özdeğerler ve özvektörler hesaplanır. Hesaplanan bu özvektörler yüksek değerden küçük değere doğru sıralanarak özellik matrisi elde edilir. Bu sıralamadaki amaç, değişkenleri, veri setini temsil kapasitesine göre sıralamaktır. Son adımda ise bu sıralama doğrultusunda istenilen sayıda alınan temsil değeri en yüksek olan bileşenler seçilerek elde edilen matris ile veri seti çarpılır ve boyutu düşürülmüş veri seti elde edilir.

3.7.9 Oto Kodlayıcı

İlk kısımlarda da bahsedildiği gibi oto kodlayıcı yapay sinir ağı modelinden türetilmiş denetimsiz bir makine öğrenmesi yöntemidir. Klasik bir yapay sinir ağı modeli üç katmandan meydana gelmektedir. Bunlar giriş katmanı, gizli katman ve çıkış katmanıdır. Her bir

katmanda belirlenen sayılarda nöronlar bulunmaktadır. Girdi katmanındaki nöron sayısını veri setindeki özellik sayısı belirlemektedir. Çıktı katmanındaki nöron sayısını ise elde edilmesi istenen sınıf sayısı belirlemektedir. Gizli katman sayısı ve buradaki nöron sayıları ise sabit değildir ve genel olarak deneme yanılma yöntemi ile bulunmaktadır. Yapay sinir ağları çok karmaşık olmayan veri setlerinde oldukça iyi sonuçlar vermektedir ancak daha karmaşık veri setleri için yeterli başarı oranı elde edilmemektedir. Autoenconder modeli ise şekil 3 de gösterildiği gibi daha karmaşık veri setlerinde başarı oranını artırmak için benzer bir ağ yapısını iyileştirerek kullanmıştır. Oto kodlayıcı modelinde girdi katmanındaki nöron sayısı genellikle gizli katmandaki nöron sayısından daha fazladır. Bu modeli ileri beslemeli yapay sinir ağlarından ayıran en önemli ikinci özellik ise giriş veri setiyle çıkış veri setinin aynı olması dolayısıyla çıktı katmanındaki nöron sayısının girdi katmanındaki nöron sayısına eşit olmasıdır.



Şekil 3. Oto kodlayıcı mimarisi

Teknik olarak oto kodlayıcı bir sınıflama işlemi yapmaz. Temel amacı N boyutlu bir özellik vektörünü daha küçük bir boyutlu vektöre en az kayıpla düşürmektir. Bunun için öncelikle girdi katmanında tüm özellikler okunur. Daha sonra bu bilgiler eşitlik 11'de gösterildiği gibi gizli katmana aktarılır. Bu eşitlikte x_j girdi katmanındaki j'ninci nöronun değerini, y_i gizli katmandaki i'ninci nörona aktarılan değeri, n girdi katmanındaki nöron sayısını, w_{ji} girdi katmanındaki j'ninci nörondan gizli katmandaki i'ninci nörona giden ağırlığı, f ise aktivasyon fonksiyonunu (örneğin: gauss, softmax, sigmoid) temsil etmektedir.

$$y_i = f\left(\sum_{j=1}^{n} x_j \times w_{ji}\right) \qquad (11)$$

İlk aşamadan sonra elde edilen değerler eşitlik 2'de gösterildiği gibi çıktı katmanına aktarılarak, son değerler hesaplanır. Eşitlik 12'de çıktı katmanındaki j'ninci i nöronu, y_i gizli katmanındaki i'ninci nöronu, w_{ji} gizli katmanındaki i'ninci nörondan çıktı katmanındaki j'ninci nörona giden ağırlığı, m gizli katmandaki nöron sayısını, f ise aktivasyon fonksiyonunu temsil etmektedir.

$$x_j^{'} = f(\cdot) \tag{12}$$

Oto kodlayıcı modelinde temel amaç, ilk iki aşama sonrasında üretilen değerinin, girdi katmanındaki değerine benzer bir değer gelmesidir. Bu iki değerin birbirine yakın çıkması için ağırlıklar geri yayılım algoritması yardımıyla hesaplanarak sürekli güncellenir. Geri yayılım algoritması eşitlik 13'te verildiği gibi iki değer arasındaki farkın karesini minimize etmektedir.

$$min \sum_{i=1}^{n} (x_j^{'} - x_j)^2 \tag{13}$$

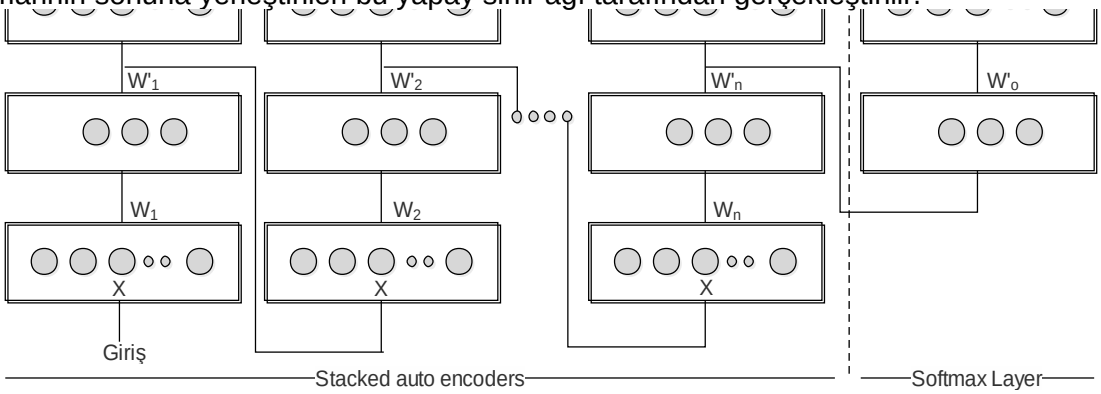
Tüm derin öğrenme mimarilerinde olduğu gibi oto kodlayıcı modelinde de aşılması gereken en önemli problemlerden birisi ağın aşırı uyum davranışına takılmasıdır (overfit). Aşırı uyum durumu, bir modelin eğitim seti için büyük başarı oranı vermesine rağmen eğitim seti dışındaki yeni bir veri seti için düşük bir başarı oranı elde etmesidir. Bu problemin önüne geçmek için çeşitli yöntemler geliştirilmiştir. Bu yöntemlerden ilki iterasyon sayısına limit koymaktır. Öğrenme aşamasında iterasyon sayısı artıkça ağırlıklar eğitim veri setini ezberleyecek derece sıfır hatayla öğrenmektedir. Giriş veri setinde gürültü olması durumunda gürültünün de gerçek veri gibi modele dâhil edilmesine neden olacaktır. Bir diğer yöntem ise, eğitim verisinin bir kısmını doğrulama verisi olarak kullanıp eğitim verilerindeki hata oranı azalırken benzer şekilde geçerlilik verisindeki hata oranın azalıp azalmadığını kontrol ederek eğitimin erken sonlandırılmasını sağlamaktır.

Aşırı uyumu engellemenin diğer yolu ise regülarizasyon yönteminin kullanılmasıdır. Bu yöntemde ise eşitlik 14'te gösterildiği geri yayılım algoritması ile güncellenecek hata değerine, ağırlıkların normları uygun bir değer ile çarpılarak eklenir. Bu yöntem sayesinde aşırı öğrenmeye neden olan kesin değişimler esnetilmektedir. Bu esnekliği ise belirlenmiş olan regülarizasyon parametresi belirlemektedir. Birçok uygulamada bu parametre deneme yanılma yöntemi ile belirlenmekte olup çok büyük ya da çok küçük olmaması gerekmektedir. Eşitlik 14'te regülarizasyon yönteminin nasıl kullanıldığı formülize edilmiştir. Bu eşitlikte toplam sembolü içerisinde ki ilk kısım eşitlik 13'ü, $L(w)$ ağırlıkların normlarını, λise regülarizasyon parametresini temsil etmektedir.

$$min¿] \quad (14)$$

Şekil 4'te gösterildiği gibi oto kodlayıcıların yığın (stacked) şeklinde arka arkaya bağlanmasıyla derin öğrenme mimarisi elde edilir. Oto kodlayıcılardan farklı olarak bu mimarinin son katmanına denetimli öğrenme işlemini gerçekleştirecek softmax içeren bir katman eklenmesiyle sınıflama ya da tahmin işlemleri gerçekleştirilir.

Arka arkaya bağlanan oto kodlayıcılar veri setinde boyut düşürme işlemlerini kademeli olarak gerçekleştirirler. Mimaride, her bir oto kodlayıcının çıkışı, bir sonraki oto kodlayıcı girişine bağlıdır. En son kodlayıcının çıkışı ise denetimli öğrenmeyi gerçekleştirecek yapay sinir ağının giriş verisini oluşturur. Bu ağın çıkışındaki nöron sayısı ise veri setindeki sınıf sayısı kadardır. Son katmandaki bu ağa kadar oto kodlayıcılarda sadece boyut düşürme işlemi gerçekleştirilir ve her hangi bir denetimli öğrenme söz konusu değildir. Sınıflama işlemi mimarinin sonuna yerleştirilen bu yapay sinir ağı tarafından gerçekleştirilir.



Şekil 4. Yığın şeklinde bağlı oto kodlayıcılı derin öğrenme makinası.

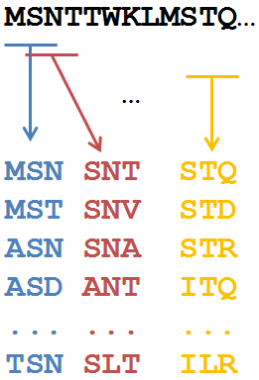
Yığın şeklinde bağlanmış oto kodlayıcılar ve son katmandaki ağ birbirinden bağımsız şekilde eğitilirler. Bu durum sistemin başarı oranı istenilen düzeye ulaştıracak yeterlilikte değildir. Sistemin daha yüksek başarımını için hassas ayarlama (fine tuning) adı verilen bir yöntem kullanılmaktadır. Bu yöntemde mimari bir bütün olarak ele alınıp ağırlık güncellenmeleri tüm sistem üzerinden yeniden gerçekleştirilir. Hassas ayarlama yöntemi detayları Yasin Görmez'in tezinde açıklanmıştır.

3.8 Parçacık seçimi

Parçacık seçimi hedef proteinin belirli bir alt dizisi için bir kütüphaneden yapısı bilinen amino asit parçalarının seçilmesidir (Simons et al., 1997). Bunun için yaygın olarak hedef protein üzerinde soldan sağa doğru kayan ve birbiriyle örtüşen pencereler alınır ve her pencereye karşılık gelen amino asit dizisine yapısal olarak benzeme ihtimali olan yüzlerce parçacık seçilir (Şekil 5). Bu pencerelerin uzunluğu değişmekle beraber 1 ila 20 amino asit

uzunluğunda olabilmektedir. Parçacık seçimi yapıldıktan sonra Monte Carlo gibi istatistiksel örnekleme algoritmaları kullanılarak pencerelerde parçacık yapıları yinelemeli olarak örneklenir ve seçilen parçacıkların birleştirilmesiyle üç boyutlu yapı modelleri kurularak en yüksek skora (en düşük enerji) sahip olan modeller belirlenir.

Projede izlenen parçacık seçimi yaklaşımında hedef protein üzerinde kayan bir pencere alınır ve her bir penceredeki amino asit dizisiyle ile kütüphanedeki parçacık yapıların dizisel ya da profil eşleşmeleri skorlanarak sıralanır



Şekil 5. Üç amino asitlik pencerelerde protein parçacık seçimi

4. BULGULAR

4.1 Model Optimizasyon Sonuçları

Bu bölümde iki aşamalı DSPRED yönteminin ikinci aşamasındaki sınıflandırıcı için destek vektör makinesi, rastgele orman ve derin KSA (deep CNF) yöntemlerinin hiper-parametrelerinin optimize edilmiş değerleri ve doğrulama (validation) kümelerindeki başarı oranları verilmiştir. Burada eşik değeri, yapısal profil matrislerini oluşturmakta kullanılan ve PDB veritabanında bulunan taslak proteinlerin hedef proteine olan maksimum dizi benzerlik yüzdesini (percentage of sequence identity) gösterir. Örneğin hedef protein HHBlits'in ikinci aşaması ile bir PDB proteinine hizalandığında hizalama skoru eşik değerinden yüksekse bu PDB proteini yapısal profil matrisinin hesaplanmasında kullanılmaz. Eşik değeri'nin %20 olduğu durum zor tahmin kategorisini, %50 olduğu durum ise orta derece zorluğu göstermektedir.

4.1.1 DVM için Optimizasyon Sonuçları

Tablo 1 ve 2 DVM için farklı eşik değerlerindeki en iyi C ve gamma parametrelerini göstermektedir. CB513'teki çapraz doğrulama deneyinin her katı için ayrı bir optimum değer bulunmuştur. Tablo 1 ve 2'deki başarı oranları doğrulama (validasyon) kümelerinde elde edilmiştir ve bir miktar aşırı uyum davranışı sergileyebilir. Yöntemlerin asıl tahmin başarısı bir sonraki bölümde verilecektir.

Tablo 1. Destek vektör makinasının CB513 doğrulama kümelerinde optimum C ve gamma parametreleri (eşik değeri=20)

Kat	C	Gamma	Q_3
1	32	0.00195313	84.0
2	32	0.00195313	81.3
3	2	0.03125	84.3
4	32	0.00195313	83.7
	8192	0.000122070	83.7
5	2	0.0078125	84.5
6	2	0.03125	82.7
7	2048	0.000122070	83.2
		Genel Doğruluk:	**83.4**

Tablo 2. Destek vektör makinasının CB513 doğrulama kümelerinde optimum C ve gamma parametreleri (eşik değeri=50)

Kat	C	Gamma	Q_3
1	512	0.000122070	89.7
2	2048	0.000122070	88.0
3	32	0.000488281	87.9
4	8	0.0078125	85.9
5	2048	0.000122070	89.4
6	512	0.000488281	87.7
7	512	0.000122070	87.8
		Genel Doğruluk:	88.0

4.1.2 Rastgele Orman için Optimizasyon Sonuçları

Tablo 3 ve 4, rasgele orman için optimum ağaç sayısını göstermektedir. CB513'teki çapraz doğrulama deneyinin her katı için ayrı bir optimum değer bulunmuştur.

Tablo 3. Rastgele orman yönteminin CB513 doğrulama kümelerindeki optimum ağaç sayısı (eşik değeri=20)

Kat	Ağaç Sayısı	Q_3
1	375	83.0
2	500	79.9
3	425	82.8
4	500	82.5
5	225	82.8
6	300	80.9
7	100	81.9
	Genel Doğruluk:	**81.9**

Tablo 4. Rastgele orman yönteminin CB513 doğrulama kümelerindeki optimum ağaç sayısı (eşik değeri=50)

Kat	Ağaç Sayısı	Q_3
1	250	85.8
2	175	83.8
3	325	85.4
4	275	83.6
5	300	85.5
6	225	84.2
7	125	84.7
	Genel Doğruluk:	84.7

4.1.3 Derin Katmanlı Sinir Alanları için Optimizasyon Sonuçları

Bu bölümde, gizli katmanların sayısı, gizli düğüm sayısı, çekirdek penceresi genişliği ve düzenlilik (regularization) katsayısı optimize edilmiştir. Tablo 5 ve 6, derin KSA için optimum hiper parametreleri göstermektedir. Optimum düzenleme katsayısı 10 veya 50, optimum gizli düğüm sayısı 75, 100 veya 125 olarak ve optimum çekirdek pencere boyutu 3 veya 4 olduğu görülmüştür.

Tablo 5. Derin katlamalı sinir alanlarının CB513 doğrulama kümelerindeki optimum çekirdek genişliği (pencere dizisi), gizli katman sayısı, gizli düğüm sayısı (düğüm dizisi), düzenleme parametresi (eşik değeri=20).

Kat	Saklı Katman	Pencere Sözcüğü	Düğüm Sözcüğü	Regülarizasyon Katsayısı	Q_3
1	5	3,3,3,3,3	125,125,125,125,125	50	90.7
2	4	3,3,3,3	100,100,100,100	10	88.8
	3	3,3,3	100,100,100	10	88.8
3	3	3,3,3	125,125,125	50	92.0
4	5	4,4,4,4,4	125,125,125,125,125	50	89.0
5	5	3,3,3,3,3	125,125,125,125,125	50	89.8
6	5	3,3,3,3,3	75,75,75,75,75	50	91.4
7	3	4,4,4	100,100,100	100	89.6
				Genel Doğruluk:	**90.2**

Tablo 6. Derin katlamalı sinir alanlarının CB513 doğrulama kümelerindeki optimum çekirdek genişliği (pencere dizisi), gizli katman sayısı, gizli düğüm sayısı (düğüm dizisi), düzenleme parametresi (eşik değeri=50)

Kat	Saklı Katman	Pencere Sözcüğü	Düğüm Sözcüğü	Regülarizasyon Katsayısı	Q_3
1	5	3,3,3,3,3	125,125,125,125,125	100	88.3

2	3	3,3,3	100,100,100	10	86.7
3	3	3,3,3	125,125,125	10	87.4
4	5	4,4,4,4,4	125,125,125,125,125	50	85.8
5	5	3,3,3,3,3	125,125,125,125,125	50	88.5
	4	4,4,4,4	75,75,75,75	50	88.5
6	5	4,4,4,4,4	125,125,125,125,125	50	89.4
7	5	3,3,3,3,3	125,125,125,125,125	50	87.5

4.2 Bireysel Modellerin Başarı Oranları

Bu kısımda DSPRED yönteminin ikinci aşamasında kullanılan sınıflandırıcıların (destek vektör makinası, rastgele orman ve deep CNF) çapraz doğrulama sonuçları sunulmuştur. Her çapraz geçerlilik tekrarlaması için optimum hiper parametreler bulunduktan sonra, modeller eğitim setlerinin tamamı üzerinde eğitilir ve tahminler test setleri üzerinde hesaplanır. Tablo 7 ve 8, DVM metodunun CB513 üzerindeki 7 katlı çapraz doğrulama deneyinde elde edilen ikincil yapı tahmini doğruluk oranlarını göstermektedir. Tablo 10'da gösterildiği gibi, doğruluk oranları, eşik değeri 20 olduğunda, %85 doğruluğa sahip olan yedinci kat hariç, test verilerinde yaklaşık %82 ile %83 arasındadır. Tablo 8'de ise eşik değeri 50 olduğunda, genel doğruluk oranının %88 olduğunu görmekteyiz. Q_E ikincil ipliklerin recall ölçütüdür ve diğer sınıf türlerinin doğruluğundan daha düşüktür. Q_H ve Q_L sırasıyla helix ve loop sınıflarının recall ölçütleridir. Q_3 ise her üç sınıfı da içeren başarı oranıdır.

Tablo 7. Destek vektör makinasının CB513 üzerindeki 7 katlı çapraz doğrulama deneyinde elde edilen ikincil yapı tahmini doğruluk oranları (eşik değeri=20).

Kat	Q_3	Q_H	Q_E	Q_L	**PrecisionH**	**PrecisionE**	**PrecisionL**
1	82.1	85.3	72.0	85.7	89.6	84.7	75.6
2	82.2	85.0	76.5	83.4	86.0	84.2	78.6
3	82.9	84.6	75.8	85.5	88.7	84.9	77.9
4	82.7	84.1	75.7	85.0	88.3	82.5	78.7
5	82.2	84.1	78.1	82.7	88.7	78.3	79.3
6	82.5	84.9	76.7	83.5	88.5	79.5	79.7
7	85.0	87.3	78.2	85.8	91.4	84.2	79.5
Genel Doğruluk:	**82.8**	**85.1**	**76.1**	**84.5**	**88.9**	**82.6**	**78.5**

Tablo 8. Destek vektör makinasının CB513 üzerindeki 7 katlı çapraz doğrulama deneyinde elde edilen ikincil yapı tahmini doğruluk oranları (eşik değeri=50).

Genel Doğruluk	Q_H	Q_E	Q_L	PrecisionH	PrecisionE	PrecisionL	Q_3
1	89.0	84.2	87.2	90.8	89.0	83.1	87.0
2	88.4	84.0	85.0	87.3	87.5	83.8	85.8
3	87.5	82.0	89.7	91.3	90.1	82.7	87.1
4	88.5	85.1	88.4	89.9	89.4	85.3	87.7
5	86.7	84.1	86.9	90.0	84.9	84.0	86.2
6	87.8	82.7	86.3	89.2	85.4	84.0	86.0
7	89.6	86.4	88.4	92.5	88.9	84.6	88.5
Genel Doğruluk:	**88.2**	**84.0**	**87.4**	**90.3**	**87.9**	**83.9**	**86.9**

Tablo 9 ve 10, rastgele orman yönteminin CB513 üzerindeki 7 katlı çapraz doğrulama deneyinde elde edilen ikincil yapı tahmini doğruluk oranlarını göstermektedir. Bu tabloda gösterildiği gibi, genel olarak, her bir kat için elde edilen doğruluklar, yakın oranlara sahiptir. Genel doğruluk oranı, eşik değeri 20 olduğunda %81.8 iken eşik değeri 50 olduğunda bu oran %84.2 olmaktadır. Her iki eşik değeri için de rasgele orman metodu uygulanarak bulunan sonuçlar DVM metodu uygulanarak bulunan sonuçlara kıyasla daha düşüktür.

Tablo 9. Rastgele orman yönteminin CB513 üzerindeki 7 katlı çapraz doğrulama deneyinde elde edilen ikincil yapı tahmini doğruluk oranları (eşik değeri=20).

Kat	Q_3	Q_H	Q_E	Q_L	PrecisionH	PrecisionE	PrecisionL
1	81.1	83.2	70.5	86.0	89.8	84.2	73.9
2	81.4	82.4	75.2	84.3	87.4	82.6	77.3
3	81.8	82.2	72.7	86.4	89.9	83.9	75.5
4	81.5	82.8	73.3	84.8	87.6	81.8	77.3
5	81.2	81.9	76.1	83.2	88.7	78.5	77.3
6	82.0	82.8	74.5	85.1	89.8	79.2	78.2
7	83.4	86.0	73.7	85.4	90.5	83.6	77.2
Genel Doğruluk:	**81.8**	**83.2**	**73.7**	**85.0**	**89.2**	**81.9**	**76.7**

Tablo 10. Rastgele orman yönteminin CB513 üzerindeki 7 katlı çapraz doğrulama deneyinde elde edilen ikincil yapı tahmini doğruluk oranları (eşik değeri=50).

Kat	Q_3

1	84.1
2	83.5
3	84.2
4	84.3
5	83.6
6	84.0
7	85.3
Genel Doğruluk:	**84.2**

Tablo 11, derin KSA yönteminin CB513 üzerindeki 7 katlı çapraz doğrulama deneyinde elde edilen ikincil yapı tahmini doğruluk oranlarını göstermektedir. Hiper parametreler için optimum değerler kullanılmıştır. Genel doğruluk oranı olarak %82.6 elde edilmiştir. Bu yöntemin eşik değeri 50 için deneyleri devam etmektedir ve yakın zamanda sonuçların elde edilmesi planlanmaktadır.

Tablo 11. Derin katlamalı sinir alanlarının CB513 üzerindeki 7 katlı çapraz doğrulama deneyinde elde edilen ikincil yapı tahmini doğruluk oranları (eşik değeri=20).

Kat	**Saklı Katman**	Q_3	Q_H	Q_E	Q_L	**PrecisionH**	**PrecisionE**	**PrecisionL**
1	5	82.1	86.2	73.1	84.4	89.6	83.5	76.1
2	4	82.0	85.5	76.5	82.8	85.3	84.1	78.8
	3	81.6	86.0	76.5	81.4	84.4	83.2	78.8
3	3	82.8	84.7	75.4	85.4	88.5	84.6	78.0
4	5	82.0	86.8	76.3	81.3	85.3	80.3	80.3
5	5	81.9	85.5	77.7	81.3	87.3	77.9	79.9
6	5	82.7	85.7	77.0	83.2	88.1	79.2	80.4
7	3	84.5	87.7	77.2	85.0	90.6	83.6	79.4
Genel Doğruluk:		**82.6**	**86.1**	**76.2**	**83.3**	**87.9**	**81.8**	**79.0**

4.3 Topluluk Metodu Sonuçları

Tablo 12, destek vektör makinası, rastgele orman ve derin KSA yöntemlerini farklı kombinasyonlarda model ortalama yaklaşımı ile birleştiren topluluk yönteminin CB513 veri

kümesindeki 7 katlı çapraz doğrulama deneyi sonuçlarını göstermektedir. Bu sonuçlara dayanarak, DVM ve derin KSA yöntemleri birleştirildiğinde en iyi doğruluk elde edilmiştir. DVM, derin KSA ve rasgele ormanı bir araya getiren topluluğun doğruluğu da benzerdir.

Tablo 12. Topluluk modellerinin CB513 üzerindeki 7 kat çapraz doğrulama deneyinde elde edilen ikincil yapı tahmini doğruluk oranları (eşik değeri=20).

Topluluk Metodu	Q_3	Q_H	Q_E	Q_L	PrecisionH	PrecisionE	PrecisionL
DVM+RO	**82.8**	84.5	75.4	85.2	89.3	83.0	78.1
DVM+DerinKSA	**83.0**	85.7	76.5	84.3	88.7	82.6	79.0
RO+DerinKSA	**82.8**	85.2	75.5	84.6	88.9	82.4	78.5
DVM+RO+DerinKSA	**83.0**	85.3	76.0	84.8	89.1	82.7	78.6

4.4 Çok Katmanlı Yapay Sinir Ağı Sonuçları

İkinci aşamadaki sınıflandırıcı için çok katmanlı perseptron (multi-layer perceptron, MLP) modeli de gerçeklenmiş, hiper-parametreleri optimize edilmiş ve CB513 veri kümesinde 7 katlı çapraz doğrulama deneyi yapılmıştır. Optimize edilen parametreler olarak saklı katman sayısı (number of hidden layers), saklı katmanlardaki düğüm sayısı (number of hidden units), öğrenme hızı (learning rate), momentum, L1 ve L2 norm katsayı değeri, dropout katsayı değeri, epoch sayısı seçilmiştir. Saklı katman sayısı 1'den 5'e adar değerler alınmıştır. Dolayısıyla derin modeller de sınanmıştır. Regülarizasyon olarak elastic net ve dropout teknikleri kullanılmıştır. Elde edilen tahmin başarı oranları %78-79 civarındadır ve DVM, derin KSA ve rastgele orman yöntemlerinden daha düşüktür. Bu deney sonucunda MLP modelinin derinleştirmenin tahmin başarı oranını iyileştirmediği sonucuna varılmıştır.

4.5 Boyut Düşürme Sonuçları

Projede uygun öznitelik boyutunun tespit edilmesi için 7 farklı öznitelik seçme ve 2 farklı izdüşüm tabanlı boyut düşürme algoritması CB513 ve EVAset olmak üzere iki farklı algoritma üstünde denenmiştir. Bu doğrultuda cross validation yöntemi kullanılarak CB513 ile 7 adet EVAset ile 10 adet veri seti elde edilmiştir. Daha sonra her bir veri seti eğitim, test, optimizasyon için eğitim ve validasyon olmak üzere 4 parçaya bölünmüştür. Her bir öznitelik seçme ve boyut düşürme algoritması için en iyi boyutu tespit etmek amacı ile optimizasyon için eğitim ve validasyon veri setleri kullanılarak DVM modelleri eğitilmiş ve validasyon veri seti için en iyi sonucu veren boyut belirlenmiştir. Daha sonra eğitim ve test veri setleri her bir yöntem için belirlenen uygun boyuta indirgenmiş ve elde edilen yeni veri setleri ile tekrar DVM modeli elde eğitilmiştir. Her bir yöntem için elde edilen ortalama boyut, katlar (fold) arası boyut farklılıkları için standart sapma değeri, precision, recall ve Q_3 başarı oranı CB513 veri seti için Tablo 13'te EVAset için Tablo 14'te gösterilmiştir.

Tablo 13. CB513 üzerinde 7-katlı çapraz doğrulama deneyi ile boyut düşürme yöntemlerinin analiz sonuçları

	Ortalama Boyut	Standart Sapma	Recall 'L'	Recall 'H'	Recall 'E'	Precision 'L'	Precision 'H'	Precision 'E'	Q3 Başarı Oranı
Orijinal Boyut	539	0	0.829	0.710	0.852	0.888	0.824	0.756	0.812
Ki-Kare	240.43	222.43	0.830	0.720	0.848	0.884	0.821	0.760	0.813
Bilgi Kazancı	225.43	203.29	0.830	0.720	0.849	0.885	0.822	0.760	0.813
Kazanım Oranı	208.43	217.97	0.830	0.720	0.850	0.886	0.823	0.761	0.814
CFS MRMR	15.43	0.79	0.828	0.716	0.841	0.876	0.816	0.758	0.808
CFS Genetik algoritma	237.43	11.09	0.827	0.713	0.852	0.888	0.822	0.757	0.812
CFS Aç Gözlü	15.43	0.79	0.828	0.716	0.841	0.876	0.816	0.758	0.808
CFS En İyi İlk Önce	15.71	0.49	0.839	0.716	0.841	0.876	0.815	0.759	0.808
TBA	120.71	75.41	0.856	0.743	0.852	0.897	0.832	0.779	0.813
Oto kodlayıcı	242.86	37.40	0.832	0.675	0.869	0.888	0.849	0.755	0.820

Tablo 14. EVAset üzerinde 10-katlı çapraz doğrulama deneyi ile boyut düşürme yöntemlerinin analiz sonuçları

	Ortalama Boyut	Standart Sapma	Recall 'L'	Recall 'H'	Recall 'E'	Precision 'L'	Precision 'H'	Precision 'E'	Q3 Başarı Oranı
Orijinal Boyut	931	0	0.865	0.780	0.845	0.894	0.837	0.795	0.838
Ki-Kare	243.90	143.46	0.858	0.767	0.846	0.892	0.835	0.786	0.833
Bilgi Kazancı	248.60	141.71	0.858	0.767	0.845	0.892	0.835	0.786	0.833
Kazanım Oranı	231.80	143.34	0.857	0.766	0.845	0.892	0.835	0.786	0.833
CFS MRMR	25.20	1.32	0.837	0.758	0.849	0.897	0.828	0.771	0.825
CFS Genetik algoritma	434.80	15.66	0.859	0.774	0.845	0.894	0.835	0.790	0.835

CFS Aç Gözlü	29.8	2.201	0.839	0.759	0.849	0.896	0.829	0.773	0.826
CFS En İyi İlk Önce	32	1.89	0.840	0.759	0.849	0.896	0.829	0.773	0.826
TBA	527	73.79	0.861	0.765	0.843	0.893	0.828	0.787	0.833
Oto kodlayıcı	427.5	24.86	0.837	0.733	0.855	0.893	0.836	0.762	0.825

Bu sonuçlara göre CB513 veri kümesinde en başarılı boyut düşürme yaklaşımı derin oto kodlayıcı ile elde edilirken, EVAset veri kümesinde genetik arama kullanan CFS (correlation feature subset evaluator) yöntemi en başarılı sonucu vermiştir. CB513 veri kümesinde boyut düşürme yaklaşımı ile tahmin başarı oranı %0.8 iyileşirken EVAset veri kümesinde %0.5 kötüleşmiştir. Bu değerler boyut düşürmenin tahmin başarı oranını fazla etkilemediğini göstermektedir. Diğer taraftan öznitelik sayısı önemli oranda azaltılmıştır. Bu sayede özellikle destek vektör makinası ve derin KSA yöntemlerinin model eğitme zamanlarının kısaltılması mümkün olacaktır.

4.6. Parçacık Seçimi

Projede parçacık seçimi için geçtiğimiz dönemlerde konsept hiyerarşi yaklaşımı ile 3-mer ve 9-mer problemi için en uygun öznitelik kombinasyonları tespit edilmişti ve logistik regresyon yöntemi eğitilmişti. Bunlar arasında 3-mer için en iyi logistik regresyon başarısını veren 75 öznitelikli kombinasyon (60 psiblast fark verisi, 9 ikincil yapı fark verisi, 3 dihedral (torsion) açı sınıfı fark verisi ve 3 çözücü erişilirlik fark verisi) 9-mer içinse en başarılı iki kombinasyon olan 9 öznitelikli kombinasyon (9 torsion fark verisi) ile 41 öznitelikli kombinasyon (20 psiblast fark verisi, 3 ikincil yapı fark verisi, 9 torsion fark verisi ve 9 çözücü erişilirlik fark verisi) seçilmiştir. Fark verilerinin nasıl hesaplandığı bir önceki proje gelişme raporunda detaylı olarak açıklanmıştı. Son dönemde ise en başarılı öznitelik kombinasyonlarını içeren veri kümesi üzerinde çeşitli makine öğrenmesi modelleri optimize edilmiş ve eğitilmiştir. Bunlar arasında Adaboost, bagging (MP5 baz öğrenicili), naive Bayes, Bayes ağı, karar ağacı (J48), en yakın komşu, rastgele orman, destek vektör makinası ve yapay sinir ağı bulunmaktadır. Destek vektör makinası libSVM programı ile, diğer yöntemler WEKA programı ile gerçeklenmiştir. Bu sınıflandırıcılar verilen iki parçacığın yapılarının birbirine benzeyip benzemediğini tahmin etmek için geliştirilmiştir. Modellerin karşılaştırılması için 10 katlı çapraz doğrulama deneyleri yapılmıştır. Model hiper-parametrelerinin optimizasyonu için her eğitim kümesinden rastgele örnekleme ile validasyon kümeleri seçilerek geriye kalan eğitim verilerinde modeller eğitilmiş ve validasyon kümeleri üzerindeki tahmin başarısını en iyileyen parametre kombinasyonu ızgara taraması (grid search) ile bulunmuştur. Aşağıdaki tabloda 3-mer problemi için geliştirilen 75 öznitelikli veri kümesi üzerinde 10 katlı çapraz doğrulama

deneyi ile optimum hiper-parametre konfigürasyonlarında elde edilen tahmin başarı ölçütleri bulunmaktadır. Bu verilerde psiblast dizi profilleri farklarının ortalaması, ikincil yapı tahmin farkı ortalaması, dihedral (torsion) açı sınıfı tahmin farkı ortalaması ve çözücü erişilirlik tahmin farkı ortalaması bulunmaktadır. Yöntemlerin başarı oranları genel olarak birbirine yakındır. En başarılı AUC (ROC eğrisi altında kalan alan) skoru bagging yöntemi ile %98.81 olarak elde edilirken, en başarılı doğruluk oranı ise rastgele orman yöntemi ile %95.12 olarak elde edilmiştir.

Tablo 15. 3-mer için 75 öznitelikli veri kümesinde 10 katlı çapraz doğrulama deney sonuçları

	NPV	Prec	Spec	Sens	MCC0	MCC1	Fm	AUC	Acc
Logistic	94.57	94.74	94.01	95.24	---	---	94.99	98.45	94.66
Adaboost	95.04	93.31	92.23	95.76	0.88	0.88	94.52	98.38	94.10
Bagging	94.76	96.28	95.83	95.32	0.91	0.91	95.80	98.81	95.56
BayesNet	93.53	93.02	91.98	94.39	0.86	0.86	93.70	95.80	93.26
J48	93.57	94.83	94.18	94.28	0.88	0.88	94.56	92.87	94.24
Knn	95.42	86.53	82.99	96.48	0.81	0.81	91.24	94.98	90.16
NBayes	93.90	91.49	90.00	94.84	0.85	0.85	93.13	94.31	92.57
RF	95.19	95.05	94.35	95.79	0.90	0.90	95.42	98.52	95.12
SVM	94.63	95.22	94.59	95.26	0.90	0.90	95.24	-	94.94

9-mer için en başarılı iki veri kümesi kullanılmıştır. Bunlardan ilkinde 9 öznitelik bulunur ve sadece dihedral açı sınıfı tahmini fark verilerinin ortalamasını (iki parçacıktan gelen verilerin farklarının ortalaması) içerir. Bu veriler üzerinde alınan 10 katlı çapraz doğrulama deneyi sonuçları aşağıdaki tabloda verilmiştir. Bu sonuçlara göre en başarılı doğruluk oranı %97.20 ile en yakın komşu yöntemi ile ve en başarılı AUC skoru ise %99.11 olarak bagging yöntemi ile elde edilmiştir.

Tablo 16. 9-mer için 9 öznitelikli veri kümesinde 10 katlı çapraz doğrulama deney sonuçları

	NPV	Prec	Spec	Sens	MCC0	MCC1	Fm	AUC	Acc
Logistic	97.15	96.34	95.80	97.52	---	---	96.92	98.71	96.71
Adaboost	96.57	96.44	95.95	96.99	0.93	0.93	96.72	98.85	96.50
Bagging	96.54	97.70	97.42	96.91	0.94	0.94	97.31	99.11	97.15
BayesNet	96.17	95.05	94.29	96.69	0.91	0.91	95.86	97.05	95.56
J48	96.20	97.82	97.56	96.60	0.94	0.94	97.20	98.19	97.05
Knn	96.74	97.61	97.31	97.11	0.94	0.94	97.36	98.89	97.20
NBayes	96.83	96.42	95.91	97.23	0.93	0.93	96.83	97.06	96.61
RF	96.48	97.57	97.26	96.86	0.94	0.94	97.21	98.64	97.05
SVM	96.75	97.45	97.12	97.12	0.94	0.94	97.28	-	97.12

Diğer 9-mer veri kümesinde ise 41 öznitelik bulunmaktadır ve psiblast dizi profilleri farklarının ortalaması, ikincil yapı tahmin farkı ortalaması, dihedral (torsion) açı sınıfı tahmin farkı ortalaması ve çözücü erişilirlik tahmin farkı ortalamasını içerir. Bu veriler üzerinde alınan 10 katlı çapraz doğrulama deneyi sonuçları aşağıdaki tabloda verilmiştir. Bu sonuçlara göre en

başarılı doğruluk oranı %96.99 ile rastgele orman ile ve en başarılı AUC skoru ise %99.16 ile bagging yöntemi ile elde edilmiştir. Rastgele orman, bagging ve en yakın komşu yöntemleri en başarılı yöntemler olarak öne çıksa da diğer yöntemlerin başarı oranları da bunlara yakındır.

Tablo 17. 9-mer 41 için öznitelikli veri kümesinde 10 katlı çapraz doğrulama deney sonuçları

	NPV	Prec	Spec	Sens	MCC 0	MCC 1	Fm	AUC	Acc
Logistic	96.10	96.65	96.21	96.55	---	---	96.60	98.96	96.39
Adaboost	95.88	96.22	95.71	96.37	0.92	0.92	96.29	98.92	96.06
Bagging	95.92	97.84	97.59	96.34	0.94	0.94	97.08	99.16	96.92
BayesNet	94.94	97.73	97.49	95.41	0.93	0.93	96.56	97.66	96.39
J48	95.08	96.50	96.07	95.61	0.92	0.92	96.05	95.87	95.83
Knn	96.94	96.60	96.12	97.32	0.93	0.93	96.96	98.62	96.75
NBayes	94.89	97.55	97.28	95.37	0.93	0.93	96.45	97.37	96.27
RF	96.19	97.71	97.43	96.60	0.94	0.94	97.15	98.89	96.99
SVM	95.92	96.63	96.20	96.39	0.93	0.93	96.51	-	96.30

5. TARTIŞMA/SONUÇ

Projede geliştirilen yöntemler neticesinde proje yürütücüsünün daha önce geliştirdiği iki aşamalı sınıflandırıcının ikincil yapı tahmin başarı oranı en zor kategoride %2.6 iyileşmiştir. Dihedral açı sınıfı tahmin başarısında da önemli oranda iyileşme elde edilmiştir. Bu iyileşme esasen yapısal profil matrislerinin iki aşamalı sınıflandırıcıya dahil edilmesinden kaynaklanmıştır. Ayrıca en zor tahmin kategorisinde eğitilen sınıflandırıcılardan üretilen tahminlerin birleştirilmesi ile doğruluk oranı biraz daha iyileşmiştir. Bu iyileşmenin yüksek eşik değerlerinde daha çok olması beklenmektedir. Burada yöntemlerin birleştirilmesi için model ortalaması yerine yığınlama gibi yaklaşımların kullanılmasıyla başarı oranındaki iyileşme daha yüksek olabilir. Yığınlama yönteminin gerçeklenmesi gelecekte yapılacak çalışmalar arasındadır.

En zor tahmin kategorisine ek olarak orta zorluk kategorisinde de model optimizasyonları yapılmıştır. Bu projenin devamında farklı zorluk seviyeleri için de model optimizasyonları tekrarlanacaktır. Bunun sonucunda verilen bir hedef proteine eşleşen PDB proteinlerinin maksimum benzerlik skoru neyse o skor için eğitilmiş (özelleşmiş) model kullanılacak ve daha başarılı bir boyutlu yapı tahminleri elde edilecektir. Bu da üç boyutlu yapı tahmin başarısını olumlu yönde etkileyecektir. Model optimizasyonu çalışmalarına ek olarak tasarlanan yeni özniteliklerin ve derin öğrenme gibi literatürdeki güncel tekniklerin yöntemlere dahil edilmesi çalışmaları devam etmektedir.

Parçacık seçimi için geliştirilen sınıflandırıcının başarı oranı da oldukça yüksek seviyededir. Burada doğrusal olmayan sınıflandırıcıların tahmin başarısı bir doğrusal sınıflandırıcı olan logistic regresyon yöntemininki ile yakın ancak biraz daha yüksektir. Dolayısıyla kullanılan

veri kümesindeki örneklerin doğrusal sınıflandırıcılar ile büyük oranda ayrıştırılmaya müsait olduğu söylenebilirse de doğrusal olmayan sınıflandırıcılar daha başarılı olduğundan öznitelik vektörü-benzerlik ilişkileri genel olarak doğrusal olmayan niteliktedir. Bu da proje önerisindeki hipotezi desteklemektedir. Ancak aradaki başarı oranı farkı çok yüksek olmadığından doğrusal sınıflandırıcılar ile de yeterince başarılı sonuçlar alınabilir. Bunu bir sonraki aşamada üç boyutlu yapı tahmin deneylerinde test etmeyi amaçlıyoruz. Ayrıca farklı özniteliklerin (dizi profilleri ve yapısal özellik tahminlerinin) kullanılmasının birbirini tamamlayıcı nitelikte olduğu sonucuna varılmıştır. Bu çalışmaların neticesinde parçacık seçimi probleminin en zor kısmı olan sınıflandırıcı geliştirilmesi başarı ile tamamlanmıştır. Bundan sonraki aşamada verilen bir hedef proteindeki parçacıkları kayan bir pencere ile tarayan, en başarılı sınıflandırıcıyı kullanarak veritabanındaki parçacıklar ile karşılaştıran ve en yüksek benzerlik skoruna sahip 200 parçacığı listeleyen bir program hazırlanacaktır. Kolay bir aşama olduğu için bu programın önümüzdeki bir iki hafta içerisinde tamamlanması planlanmaktadır. Özet olarak projede hedeflenen iş paketlerinin çoğu tamamlanmıştır.

Önümüzdeki yakın dönemde projede geliştirilen yöntemler üç boyutlu yapı tahmini için kullanılacak ve tahmin başarısı çeşitli koşullarda incelenecektir. Bu son iş paketi projede yaşanan bazı aksaklıklar sebebiyle yetiştirilememiştir. Gecikmenin bir diğer sebebi ise problemlerin büyük veri analizi kategorisinde olması ve bu nedenle model eğitme ve optimizasyonlarının uzun sürmesidir. Proje süresince proje bütçesinden alınan iş istasyonundan, kurumumuzda bulunan diğer iş istasyonlarından, TRUBA'daki kaynaklardan ve Compecta firmasının hesaplama sisteminden istifade edilmiştir. Ancak bazı deneyler TRUBA'daki ve Compecta'daki kota ve kaynak limitlerine takıldığından sadece iş istasyonlarında yapılabilmiştir. Yöntemlerin bir internet sunucusu ve internet sayfası üzerinden dünya genelinde erişme açılması da yakın gelecekte hedeflenen çalışmalar arasındadır. Tüm bu gayretler neticesinde üç boyutlu yapı tahminin iyileştirilmesi ve ilaç tasarımı çalışmalarının başarı oranlarının arttırılması arzulanmaktadır.

REFERANSLAR

Adamczak, R. 2009. "Dimensionality reduction of PSSM matrix and its influence on secondary structure and relative solvent accessibility predictions", Presented at the World Academy of Science, Engineering and Technology 58.

Alirezaee, M., Dehzangi, A., Mansoori, E. 2012. "Ensemble of neural networks to solve class imbalance problem of protein secondary structure prediction", International Journal of Artificial Intelligence & Applications, 3(6), 9.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. 1997. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic acids research, 25(17), 3389-3402.

Aydin, Z., Altunbasak, Y., Borodovsky, M. 2006. "Protein secondary structure prediction for a single-sequence using hidden semi-Markov models", BMC Bioinformatics 7, 178.

Aydin, Z., Singh, A., Bilmes, J., Noble, W. S. 2011. “Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure”, BMC bioinformatics, 12(1), 154.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., Tasumi, M. 1977. “The protein data bank: a computer-based archival file for macromolecular structures”, J Mol. Biol. , 112(3), 535-542.

Bhola, A., Yadav, S. K., Tiwari, A. K. 2014. “Machine Learning based Approach for Protein function prediction using sequence derived properties”, International journal of computer applications, 105(12).

Bologna, G., Appel, R. D. 2002. “A comparison study on protein fold recognition”, In Proceedings of the IEEE 9th International Conference on Neural Information Processing 5, 2492-2496.

Chen, K. E., Kurgan, L. A., Ruan, J. 2008. "Prediction of protein structural class using novel evolutionary collocation-based sequence representation", Journal of computational chemistry, 29(10), 1596-1604.

Cheng, J., Eickholt, J., Wang, Z., Deng, X. 2012. "Recursive protein modeling: a divide and conquer strategy for protein structure prediction and its case study in casp9", J. Bioinform. Comput. Biol. 10, 1242003.

Chinnasamy, A., Sung, W. K., Mittal, A. 2005. “Protein structure and fold prediction using tree-augmented naive Bayesian classifier”, Journal of Bioinformatics and Computational Biology, 3(04), 803-819.

Dale, J. M., Popescu, L., Karp, P. D. 2010. “Machine learning methods for metabolic pathway prediction”, BMC bioinformatics, 11(1), 15.

Dietterich, T. G. 2000. “Ensemble methods in machine learning”, Multiple classifier systems, 1857, 1-15.

Ding, C. H., Dubchak, I. 2001. “Multi-class protein fold recognition using support vector machines and neural networks”, Bioinformatics, 17(4), 349-358.

Fujitsuka, Y., Chikenji, G., Takada, S. 2006. "SimFold energy function for de novo protein structure prediction: Consensus with Rosetta", Proteins Struct. Funct. Bioinforma. 62, 381–398.

Graphical Models Toolkit (GMTK): http://melodi.ee.washington.edu/~bilmes/gmtk/.

Ghosh, A., Parai, B. 2008. "Protein secondary structure prediction using distance based classifiers", Int. J. Approx. Reason., 47(1), 37–44.

Gront, D., Kulp, D. W., Vernon, R.M., Strauss, C. E. M., Baker, D. 2011. "Generalized Fragment Picking in Rosetta: Design, Protocols and Applications", PLOS ONE 6, e23294.

Gubbi, J., Lai, D. T., Palaniswami, M., Parker, M. 2006. "Protein secondary structure prediction using support vector machines and a new feature representation", International Journal of Computational Intelligence and Applications, 6(04), 551-567.

Hua, S., Sun, Z. 2001. "A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach", J. Mol. Biol. 308, 397–407.

Huang, Y. F., Chen, S. Y. 2013. "Protein secondary structure prediction based on physicochemical features and PSSM by SVM", in: 2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). Presented at the 2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 9–15.

Jian-wei, L., Guang-hui, C., Hai-en, L., Yuan, L., Xiong-lin, L. 2013. "Prediction of protein secondary structure using multilayer feed-forward neural networks", in 25th Chinese Control and Decision Conference (CCDC), 1346–1351.

Jones, D. T. 1999. "Protein secondary structure prediction based on position-specific scoring matrices", Journal of Molecular Biology, 292, 195-202.

Jones, D. T. 2001. “Protein structure prediction in genomics.”, Briefings in bioinformatics, 2(2), 111-125.

Kabsch, W., Sander, C. 1983. “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers”, 22(12), 2577-2637.

Kavzoğlu, T., Şahin, E. K., Çölkesen, İ., 2014. "Heyelan Duyarlılık Analizinde Ki-Kare Testine Dayalı Faktör Seçimi. Presented at the V. Uzaktan Algılama ve Coğrafi Bilgi Sistemleri Sempozyumu" (UZAL CBS 2014).

Kim, H., Park, H. 2003. “Protein secondary structure prediction based on an improved support vector machines approach”, Protein Engineering, 16(8), 553-560.

King, R. D., Ouali, M., Strong, A. T., Aly, A., Elmaghraby, A., Kantardzic, M., Page, D. 2000. “Is it better to combine predictions?”, Protein Engineering, 13(1), 15-19.

Kountouris, P., Agathocleous, M., Promponas, V. J., Christodoulou, G., Hadjicostas, S., Vassiliades, V., Christodoulou, C. 2012. “A comparative study on filtering protein secondary structure prediction”, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 9(3), 731-739.

Lee, J., Lee, D., Park, H., Coutsias, E. A., Seok, C. 2010. "Protein loop modeling by using fragment assembly and analytical loop closure", Proteins Struct. Funct. Bioinforma. 78, 3428–3436.

Lee, J., Kim, S.-Y., Joo, K., Kim, I., Lee, J. 2004. "Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing", Proteins Struct. Funct. Bioinforma. 56(4), 704–714.

Lee, J., Lee, J., Sasaki, T. N., Sasai, M., Seok, C., Lee, J. 2011. "De novo protein structure prediction by dynamic fragment assembly and conformational space annealing", Proteins Struct. Funct. Bioinforma. 79(8), 2403–2417.

Li, D., Li, T., Cong, P., Xiong, W., Sun, J. 2011. “A novel structural position-specific scoring matrix for the prediction of protein secondary structures”, Bioinformatics, 28(1), 32-39.

Li, S. C., Bu, D., Xu, J., Li, M. 2008. "Fragment-HMM: A new approach to protein structure prediction", Protein Sci. 17, 1925–1934.

Li, Z., Wang, J., Zhang, S., Zhang, Q., Wu, W. 2017. "A new hybrid coding for protein secondary structure prediction based on primary structure similarity", Gene 618, 8–13.

Li, Z., Yu, Y. 2016. "Protein Secondary Structure Prediction Using Cascaded Convolutional and Recurrent Neural Networks", ArXiv160407176 Cs Q-Bio.

Martin, J., Gibrat, J.-F., Rodolphe, F. 2006. "Analysis of an optimal hidden Markov model for secondary structure prediction", BMC Struct. Biol., 6(25).

Mirabello, C., Pollastri, G. 2013. "Porter, PaleAle 4.0: high accuracy prediction of protein secondary structure and relative solvent accessibility", Bioinformatics, 29(16), 2056-2058.

Perrin, B. E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., d’Alche–Buc, F. 2003. “Gene networks inference using dynamic Bayesian networks”, Bioinformatics, 19(suppl_2), ii138-ii148.

Pollastri, G., Baldi, P., Fariselli, P., Casadio, R. 2002a. "Prediction of coordination number and relative solvent accessibility in proteins", Proteins: Structure, Function, and Bioinformatics, 47(2), 142-153.

Pollastri, G., Przybylski, D., Rost, B., Baldi, P. 2002b. "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles", Proteins Struct. Funct. Bioinforma., 47(2), 228–235.

Pollastri, G., McLysaght, A. 2005. "Porter: a new accurate server for protein secondary structure prediction", Bioinformatics, 21(8), 1719-1720.

Remmert, M., Biegert, A., Hauser, A., Söding, J. 2012. "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment", Nature methods, 9(2), 173-175.

Reynolds, S. M., Käll, L., Riffle, M. E., Bilmes, J. A., Noble, W. S. 2008. "Transmembrane topology and signal peptide prediction using dynamic bayesian networks", PLoS computational biology, 4(11), e1000213.

Roy, A., Kucukural, A., Zhang, Y., 2010. "I-TASSER: a unified platform for automated protein structure and function prediction", Nat. Protoc. 5, 725–738.

Salamov, A. A., Solovyev, V. V., 1995. "Prediction of Protein Secondary Structure by Combining Nearest-neighbor Algorithms and Multiple Sequence Alignments", J. Mol. Biol. 247, 11–15.

Simoncini, D., Berenger, F., Shrestha, R., Zhang, K. Y. J. 2012. "A Probabilistic Fragment-Based Protein Structure Prediction Algorithm", PLOS ONE 7, e38799.

Simons, K. T., Kooperberg, C., Huang, E., Baker, D., 1997. "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions", J. Mol. Biol. 268, 209–225.

Spencer, M., Eickholt, J., Cheng, J. 2015. "A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction", IEEE/ACM Trans. Comput. Biol. Bioinform., 12(1), 103–112.

Tian, L., Wu, A., Cao, Y., Dong, X., Hu, Y., Jiang, T. 2011. "NCACO-score: An effective main-chain dependent scoring function for structure modeling", BMC Bioinformatics 12, 208.

Wang, Y., Cheng, J., Liu, Y., Chen, Y. 2016. "Prediction of protein secondary structure using support vector machine with PSSM profiles", in: IEEE Information Technology, Networking, Electronic and Automation Control Conference. Presented at the 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference, pp. 502–505.

Wang, S., Peng, J., Ma, J., Xu, J. 2016. "Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields", Sci. Rep., 6.

Ward, J. J., McGuffin, L. J., Buxton, B. F., Jones, D. T. 2003. “Secondary structure prediction with support vector machines.”, Bioinformatics, 19(13), 1650-1655.

Xu, D., Zhang, Y., 2012. "Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field", Proteins Struct. Funct. Bioinforma. 80, 1715–1735.

Yao, X. Q., Zhu, H., She, Z. S. 2008. “A dynamic Bayesian network approach to protein secondary structure prediction”, BMC bioinformatics, 9(1), 49.

Yang, W., Wang, K., Zuo, W. 2011. "A fast and efficient nearest neighbor method for protein secondary structure prediction", in 3rd International Conference on Advanced Computer Control, 224–227.

Zeng, J., Zhu, S., Yan, H. 2009. “Towards accurate human promoter recognition: a review of currently used sequence features and classification methods”, Briefings in bioinformatics, 10(5), 498-508.

Zhang, J., He, Z., Wang, Q., Barz, B., Kosztin, I., Shang, Y., Xu, D., 2012. "Prediction of Protein Tertiary Structures Using MUFOLD", in: Functional Genomics, Methods in Molecular Biology. Springer, New York, NY, pp. 3–13.

Zhou, T., Shu, N., Hovmöller, S. 2009. “A novel method for accurate one-dimensional protein structure prediction based on fragment matching”, Bioinformatics, 26(4), 470-477.

TÜBİTAK
PROJE ÖZET BİLGİ FORMU

Proje Yürütücüsü:	Yrd. Doç. Dr. ZAFER AYDIN
Proje No:	113E550
Proje Başlığı:	Zenginleştirilmiş Öznitelikler Ve Makine Öğrenmesi Yöntemleriyle Protein Yerel Yapı Tahmini
Proje Türü:	3501 - Kariyer
Proje Süresi:	30
Araştırmacılar:	
Danışmanlar:	
Projenin Yürütüldüğü Kuruluş ve Adresi:	ABDULLAH GÜL Ü.
Projenin Başlangıç ve Bitiş Tarihleri:	01/09/2014 - 01/09/2017
Onaylanan Bütçe:	191489.0
Harcanan Bütçe:	95393.76
Öz:	Projenin amacı proteinlerde bulunan ikincil yapı, dihedral açı ve çözücü erişilirlik gibi bir boyutlu yapısal özelliklerin başarılı olarak tahmin edilmesi ve bu tahminleri kullanarak parçacık seçimi yapan yeni bir yöntem geliştirilmesidir. Geliştirilen yöntemler sayesinde proteinlerin üç boyutlu yapısının daha doğru tahmin edilmesi, proteinlerin fonksiyonlarının daha iyi anlaşılması ve daha etkili ilaç tasarımı yapılması mümkün olacaktır. Bir boyutlu yapısal özelliklerin tahmini için yürütücünün daha önce geliştirdiği iki aşamalı hibrit sınıflandırma yöntemi kullanılmıştır. Bu yöntemde bulunan sınıflandırıcılar için dizi tabanlı profiller, yapısal profil matrisleri gibi çeşitli öznitelik vektörleri kullanılmıştır. İkinci aşamadaki sınıflandırıcı için destek vektör makinası, derin CNF, rastgele orman ve topluluk gibi çeşitli öğrenme yöntemleri eğitilmiş ve geliştirilen yöntemlerin tahmin başarı oranları standart veri kümelerinde incelenmiştir. Ayrıca bu aşamada derin otokodlayıcılar ve öznitelik seçme yaklaşımları ile boyut düşürme gerçekleştirilmiştir. Protein parçacık seçimi için verilen iki amino asit dizisi parçacığının yapısal olarak benzer olup olmadığının tahmin eden yöntemler geliştirilmiştir. Bunun için Rosetta programının parçacık veritabanında bulunan proteinlerden parçacık ikilileri örneklenmiş, bu ikililer BCScore yöntemi ile etiketlenmiş, eğitim ve test kümeleri oluşturulmuştur. Ayrıca farklı öznitelik kümeleri konsept hiyerarşi yaklaşımı ile kapsamlı olarak incelenmiş ve en başarılı sonucu veren öznitelik kombinasyonları tespit edilmiştir. Parçacık seçimi probleminde 3 ve 9 amino asitlik parçacıklar üzerinde çalışılmıştır ancak yöntemler diğer uzunluktaki parçacıklar için de kolaylıkla uygulanabilecektir. Projede geliştirilen yöntemler sayesinde ikincil yapı tahmin başarısı en zor tahmin kategorisinde %2.6 iyileşmiş, dihedral açı tahmin başarısı önemli oranda iyileşmiş, çözücü erişilirlik probleminde literatürdeki en başarılı yöntemler ile benzer bir seviye yakalanmıştır. Parçacık seçiminde ise verilen iki parçacığın yapılarının benzer olup olmadıkları 3-mer parçacıklar için %94 ve 9-merler içinse %97 oranı ile tahmin edilmiştir. Yapılan çalışmaların neticesinde öznitelik vektörlerinin daha iyi tasarlanmasının ve farklı sınıflandırma yöntemlerinin birleştirilip optimize edilmesinin yapısal özellik tahmin başarısını önemli oranda iyileştirdiği sonucuna varılmıştır.
Anahtar Kelimeler:	Protein yapı tahmini, protein parçacık seçimi, makine öğ?renmesi, öznitelik çıkarımı, boyut düşürme
Fikri Ürün Bildirim Formu Sunuldu Mu?:	Hayır
Projeden Yapılan Yayınlar:	1- Template Scoring Methods for Protein Torsion Angle Prediction (Makale - Diğer Hakemli Makale),