

**T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**İYİLEŞTİRİLMİŞ KLONAL SEÇİM ALGORİTMASI
İLE HEMOGLOBİN PROTEİNİ İKİNCİL YAPI
TAHMİNİ**

DOKTORA TEZİ

Burcu ÇARKLI YAVUZ

**Enstitü Anabilim Dalı : BİLGİSAYAR VE BİLİŞİM
MÜHENDİSLİĞİ**

Tez Danışmanı : Doç. Dr. Nilüfer YURTAY

Aralık 2018

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**İYİLEŞTİRİLMİŞ KLONAL SEÇİM ALGORİTMASI
İLE HEMOGLOBİN PROTEİNİ İKİNCİL YAPI
TAHMİNİ**

DOKTORA TEZİ

Burcu ÇARKLI YAVUZ

Enstitü Anabilim Dalı

: BİLGİSAYAR VE BİLİŞİM
MÜHENDİSLİĞİ

Bu tez 20 / 12 /2018 tarihinde aşağıdaki jüri tarafından oybirliği/oyçokluğu ile kabul edilmiştir.



**Prof. Dr.
Cemil ÖZ**
Jüri Başkanı



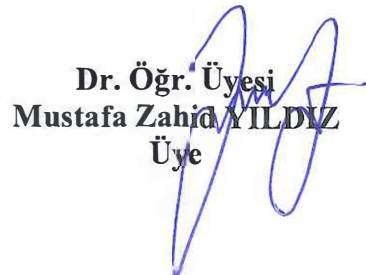
**Doç. Dr.
Nilüfer YURTAY**
Üye



**Doç. Dr.
Mehmet Recep BOZKURT**
Üye



**Doç. Dr.
Kürşat AYAN**
Üye



**Dr. Öğr. Üyesi
Mustafa Zahid YILDIZ**
Üye

BEYAN

Tez içindeki tüm verilerin akademik kurallar çerçevesinde tarafımdan elde edildiğini, görsel ve yazılı tüm bilgi ve sonuçların akademik ve etik kurallara uygun şekilde sunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, tezde yer alan verilerin bu üniversite veya başka bir üniversitede herhangi bir tez çalışmasında kullanılmadığını beyan ederim.

Burcu ÇARKLI YAVUZ

20.12.2018

TEŞEKKÜR

Yüksek lisans ve doktora eğitimim boyunca değerli bilgi ve deneyimlerinden yararlandığım, her konuda bilgi ve desteğini almaktan çekinmediğim, araştırmanın planlanmasından yazılmasına kadar tüm aşamalarında yardımlarını esirgemeyen değerli danışman hocam Doç. Dr. Nilüfer YURTAY'a teşekkürlerimi sunarım.

Tez çalışmamın şekillenmesinde büyük katkıları olan, titizlikle beni yönlendiren kıymetli hocalarım Prof. Dr. Cemil ÖZ ve Doç. Dr. Mehmet Recep BOZKURT'a teşekkür ederim. Ayrıca tez çalışmam için gerekli olan biyolojik altyapının oluşması için desteklerini esirgemeyen Prof. Dr. Hikmet Benan DİNÇTÜRK BOTOFTÉ'nin de katkıları büyüktür, kendisine teşekkür ederim.

Akademik hayatımın bana kazandırdığı en özel üç dost; Ceyda AKSOY TIRMIKÇI, Halime HIZARCI ve Özhan ÖZKAN, desteğiniz, yardımlarınız ve çalışmamdaki katkılarınız çok önemli.

En büyük şansım sevgili eşim ve canım ailem, bu uzun ve zorlu yolculuğumda her zaman yanımda olduğunuz için hepinize minnettarım, sabrınız ve fedakârlıklarınız için çok teşekkür ederim.

Bu çalışmanın maddi açıdan desteklenmesine olanak sağlayan Sakarya Üniversitesi Bilimsel Araştırma Projeleri (BAP) Komisyon Başkanlığına (Proje No: 2013-50-02-029) teşekkür ederim.

İÇİNDEKİLER

TEŞEKKÜR	i
İÇİNDEKİLER	ii
SİMGELER VE KISALTMALAR LİSTESİ	iv
ŞEKİLLER LİSTESİ	v
TABLolar LİSTESİ	vi
ÖZET	vii
SUMMARY	viii
BÖLÜM 1.	
GİRİŞ	1
BÖLÜM 2.	
LİTERATÜR ÖZETİ.....	3
BÖLÜM 3.	
PROTEİNLER	11
3.1. Proteinlerin Yapısı	13
3.1.1. Birincil yapı	13
3.1.2. İkincil yapı	17
3.1.3. Üçüncül yapı	18
3.1.4. Dördüncül yapı	19
3.2. Laboratuvar Ortamında Proteinlerin Yapısal Analizi	19
3.2.1. X-ışını kristalografisi	19
3.2.2. Nükleer manyetik rezonans spektroskopisi	20
3.2.3. Kriyoelektron mikroskobu	20

3.2.4. Atomik kuvvet mikroskobu	21
3.3. Hatalı Katlanma	21
BÖLÜM 4.	
MATERYAL VE YÖNTEM.....	23
4.1. Doğal ve Yapay Bağışıklık Sistemleri	23
4.1.1. Doğal Bağışıklık Sistemi.....	23
4.1.2. Yapay Bağışıklık Sistemi	26
4.1.2.1. Klonal seçim algoritması.....	26
4.1.2.2. Negatif seçim algoritması.....	33
4.1.2.3. Bağışıklık ağı teorisi.....	34
4.2. Sınıflandırıcılar	34
4.2.1. Çok Katmanlı Algılayıcı	35
4.2.2. K-En Yakın Komşu Algoritması	37
4.2.3. Naive Bayes Sınıflandırıcısı.....	38
4.2.4. ID3 Algoritması	39
4.3. Uygulama Mimarisi	40
BÖLÜM 5.	
UYGULAMA ÇIKTILARI	49
BÖLÜM 6.	
SONUÇLAR VE ÖNERİLER	57
KAYNAKLAR	60
ÖZGEÇMİŞ	69

SİMGELER VE KISALTMALAR LİSTESİ

Ab	: Antibody (Antikor)
Ag	: Antigen (Antijen)
CSA	: Clonal Selection Algorithm (Klonal Seçim Algoritması)
ÇKA	: Çok Katmanlı Algılayıcı
DNA	: Deoksiribonükleik Asit
DSSP	: Dictionary of Protein Secondary Structure (Protein İkincil Yapı Sözlüğü)
ID3	: Iterative Dichotomiser 3 (Tekrarlı İkilikçi Ağaç)
KSA	: Klonal Seçim Algoritması
kNN	: k Nearest Neighbor (k En Yakın Komşu)
MLP	: Multilayer Perceptron (Çok Katmanlı Algılayıcı)
mRNA	: Mesajcı RNA
NB	: Naive Bayes
NMR	: Nükleer Manyetik Rezonans
P	: Popülasyon
PDB	: Protein Data Bank (Protein Veri Bankası)
Q ₃	: 3 sınıf için tahmin başarısı
Q ₈	: 8 sınıf için tahmin başarısı
RNA	: Ribonükleik Asit
rRNA	: Ribozomal RNA
tRNA	: Taşıyıcı RNA
YBS	: Yapay Bağışıklık Sistemi

ŞEKİLLER LİSTESİ

Şekil 3.1. İyonlaşmamış haldeki genel aminoasit yapısı	11
Şekil 3.2. İyonize olmuş aminoasit yapısı.....	12
Şekil 3.3. Peptit bağ oluşumu	12
Şekil 3.4. Protein yapısında bulunan 20 aminoasit.....	13
Şekil 3.5. Proteinlerin birincil, ikincil, üçüncül ve dördüncül yapıları	14
Şekil 3.6. Protein sentezinde genetik bilginin akışı	15
Şekil 3.7. Genetik (üçlü) kod aminoasit dizilimini belirler.....	16
Şekil 3.8. Proteinlerin ikincil yapısında bulunan motifler	18
Şekil 4.1. Önerilen çalışmanın blok diyagramı.....	23
Şekil 4.2. Bağışıklık sisteminin çok katmanlı yapısı	24
Şekil 4.3. Klonal seçim prensibi	25
Şekil 4.4. Klonal seçim algoritması blok diyagramı	28
Şekil 4.5. ÇKA yapısı	35
Şekil 4.6. kNN algoritmasında k değerinin belirlenmesi	38
Şekil 4.7. 1A4F kodlu Hemoglobin proteinin birincil ve ikincil yapısı	41
Şekil 4.8. C# form ekranı	42
Şekil 4.9. Kayan pencere yöntemi	43
Şekil 4.10. Kayan pencere yöntemi uygulandığında elde edilen veri formatı	44
Şekil 4.11. Veri tabanında kullanılan tablolar ve saklı yordamlar	45
Şekil 4.12. Rapidminer ekran 1	47
Şekil 4.13. Rapidminer ekran 2	48
Şekil 5.1. Her antijen için en benzer antikör ve toplam benzerlikler	52

TABLÖLER LİSTESİ

Tablo 4.1. KSA için örnek gözlem tablosu.....	29
Tablo 4.2. Eğitim için ayrılan veri (Antikorlar).....	29
Tablo 4.3. Test için ayrılan veri (Antijenler).....	30
Tablo 4.4. Pn tablosu.....	31
Tablo 4.5. Pn klonlu tablosu.....	31
Tablo 4.6. Pn mutasyonlu tablosunun oluşturulması.....	32
Tablo 4.7. KSA ile eğitilmiş veri.....	32
Tablo 4.8. Çalışmada kullanılan proteinler.....	41
Tablo 5.1. 1A4F proteininin farklı pencere boyutları ve iterasyonlar için sınıflandırma başarı değişimi	49
Tablo 5.2. Pencere boyutlarına göre başarı değişimi	50
Tablo 5.3. KSA ve kNN ile yapılan hemoglobin ikincil yapı tahmininin farklı yöntemler ile mukayesesi	51
Tablo 5.4. Tek test ile eğitim yapıldığında başarı	52
Tablo 5.5. Her iterasyonda değişen testler ile eğitim yapıldığında başarı	52
Tablo 5.6. KSA ile 5 iterasyon yapıldıktan sonra ÇKA ile sınıflandırma başarıları	53
Tablo 5.7. KSA ile 6 iterasyon yapıldıktan sonra ÇKA ile sınıflandırma başarıları	54
Tablo 5.8. KSA ile 7 iterasyon yapıldıktan sonra ÇKA ile sınıflandırma başarıları	54
Tablo 5.9. KSA ile 8 iterasyon yapıldıktan sonra ÇKA ile sınıflandırma başarıları	55
Tablo 5.10. Nöron sayısı değişimine bağlı sınıflandırma başarıları	55
Tablo 5.11. Aktivasyon fonksiyonu değişimine bağlı sınıflandırma başarıları	56
Tablo 5.12. KSA uygulanmadan ve uygulandıktan sonra yapılan sınıflandırmaların başarı oranları.....	56

ÖZET

Anahtar kelimeler: Protein ikincil yapı tahmini, hemoglobin, klonal seçim, k en yakın komşu, çok katmanlı algılayıcı, ID3, Naive Bayes

Protein ikincil yapının biliniyor olması proteinin işlevinin belirlenmesinde, birçok hastalığın tedavi edilmesinde ve ilaç tasarımıyla çok önemli avantajlar sağlamaktadır. Laboratuvar ortamında ikincil yapı tespiti hem maliyetli hem de zorlu bir süreçtir. Bu nedenle protein ikincil yapı tahmini, uzun yıllardır biyoinformatik ve hesaplamalı biyolojinin önemli bir çalışma alanıdır.

İki aşamalı olan bu çalışmada doğadan ilham alan yöntemler kullanılarak protein ikincil yapı tahminine katkı sağlanması amaçlanmıştır. Birinci aşamada veri, canlı bağışıklık sisteminden esinlenerek modellenen Klonal Seçim Algoritması (KSA) ile eğitilmiştir. KSA’da geleneksel yöntemden farklı olarak, her bir antijene en benzer antikor seçimi yerine, toplam benzerlikleri dikkate alınarak seçim yapılmıştır. İkinci aşamada farklı yöntemler kullanılarak sınıflandırma yapılmıştır. Biyolojik sinir sisteminden esinlenerek modellenen yöntemlerden biri olan çok katmanlı algılayıcı (ÇKA), mesafeye dayalı k en yakın komşu algoritması (kNN), karar ağaçları temelli ID3 ve istatistiksel hesaplara dayalı Naive Bayes (NB) yöntemleri sınıflandırma için tercih edilmiştir.

Alınan sonuçlar KSA’da önerilen yeni antikor seçimi yaklaşımının geleneksel KSA modeline göre daha iyi sonuçlar verdiğini göstermiştir. Geliştirilen KSA modeli ile hemoglobin veri seti eğitilmiş ve eğitilen protein verisinin ikincil yapı tahmini farklı sınıflandırıcı yöntemleri ile yapılmıştır. Çalışmanın sonunda KSA ile eğitilen verinin sınıflandırma başarısının eğitilmeyen veriye göre daha yüksek olduğu görülmüştür.

HEMOGLOBIN PROTEIN SECONDARY STRUCTURE PREDICTION BY USING IMPROVED CLONAL SELECTION ALGORITHM

SUMMARY

Keywords: Protein secondary structure prediction, hemoglobin, clonal selection, k nearest neighbors, multilayer perceptron, ID3, Naive Bayes

The protein secondary structure information provides very important advantages in determining the function of a protein, treating numerous diseases and drug design. Determining the secondary structure in the laboratory environment is both costly and challenging. Therefore, the prediction of protein secondary structure has been an important study field of bioinformatics and computational biology for many years.

The aim of the present two-phased study is to provide a contribution to the prediction of protein secondary structure using the nature-inspired methods. The data in the first phase were trained with Clonal Selection Algorithm (CSA) which was modeled by being inspired by the live immune system. CSA was improved by using the total similarity for the selection instead of the traditional way of the most similar antibody for an antigen. Using different methods classification is realized in the second phase. The classification was then performed with Multilayer Perceptron (MLP) which is modeled by being inspired by the biological nervous system, k nearest neighbor (kNN) algorithm which is based on distance, decision tree based ID3 method and Naive Bayes (NB) method based on statistical calculations.

The results obtained indicated that the proposed antibody selection method performed better than the traditional method. The hemoglobin data set was trained by the new CSA model and the secondary structure prediction of the trained set was realized by different classification methods. In conclusion it was found out that the classification success of the trained data set was better than the success of the untrained data set.

BÖLÜM 1. GİRİŞ

Proteinler bütün canlı sistemlerde bulunur ve her türlü biyolojik işlemin yerine getirilmesinde görev alırlar [1]. Proteinin üç boyutlu şekli ve aminoasit yan zincirlerinin özgün kimyasal özellikleri proteinin işlevini belirler. Bunun sonucunda proteinler hücre içinde ve hücre dışında yaşamsal faaliyetler bakımından önem taşıyan veya hücre ve organizmaya evrimsel bir avantaj sağlayan bir seri önemli işlevi yerine getirirler [2].

Proteinler birincil, ikincil, üçüncül ve dördüncül yapılarda bulunabilirler. Aminoasitler peptit bağlarıyla bir araya gelerek bir proteinin birincil yapısını oluştururlar. Aminoasitlerin karboksil ve amino grupları arasında oluşan hidrojen bağları ikincil yapıyı şekillendirir. İkincil yapı, proteinlerin üç boyutlu şeklinin ve fonksiyonunun belirlenmesinde önemli bir adımdır. İkincil yapıdaki en yaygın yapısal motifler sarmal, tabaka ve iplik yapılardır. Üçüncül yapı, polipeptit zincirinin bütünsel şekline yani tüm aminoasitlerin üç boyutlu düzenlenmesine denir. Dördüncül yapı bir ya da daha fazla polipeptitten oluşmaktadır.

Polipeptit zincirdeki aminoasitler birbiri üzerine katlanarak ikincil ve üçüncül yapıyı şekillendiren bağları oluşturdıkları için bu süreç protein katlanması olarak da adlandırılır. İkincil yapı birbirine yakın aminoasitler arasındaki etkileşimlerden oluştuğu için bölgesel katlanma, üçüncül yapı ise geniş aralıklı etkileşimlerden oluştuğu için tam katlanma olarak adlandırılır.

Proteinler hücredeki her işlemde yer alır, bu nedenle organizma için hayati öneme sahiptirler. İşlevi bozulmuş proteinler yıkıcı sonuçlar doğurabilir. Yaşlandıkça mutasyonlar ve bazı dış faktörler proteinlerin yanlış katlanmasına neden olur.

Günümüzde özellikle ileri yaştaki insanlarda sıklıkla karşılaştığımız Alzheimer, Parkinson, Tip 2 diyabet ve katarakt hatalı katlanan proteinlerin neden olduğu hastalıklara örnek olarak verilebilir [3].

Günlük hayatta sıklıkla karşılaştığımız bu hastalıklar kişinin yaşam kalitesini düşürmektedir. İşin maddi boyutu ele alındığında, bu hastalıkların tedavisi için oldukça yüksek harcamaların yapıldığı da bilinmektedir. Hatalı katlanan proteinlerin tespiti birçok genetik bozukluk ve hastalık ile mücadelede avantaj sağlayacaktır. Genetik bozukluklara ve hastalıklara müdahale edebilmek ve yeni tedaviler geliştirebilmek için proteinlerin üç boyutlu yapılarının bilinmesi oldukça önemlidir.

Proteinin doğru katlanması için gerekli bilginin aminoasit dizisinde bulunduğu laboratuvar ortamında yapılan çalışmalarla kanıtlanmıştır. Denatürasyon şartlarında proteinlerin katlanmalarında bozulmalar görülmüştür, normal şartlara döndüğünde bu proteinlerin bazıları tekrar doğal şekillerini kazanmışlardır. Bu durum göstermektedir ki proteinlerin düzgün olarak katlanabilmeleri için gerekli bilgi onların birincil yapısında mevcuttur. Birincil yapı üç boyutlu yapıyı, üç boyutlu yapı ise proteinin işlevini belirler [2].

Laboratuvar ortamında proteinlerin üç boyutlu yapılarının belirlenmesi çok masraflı ve zorlu bir süreçtir. Her proteine uygulanabilen genel bir yöntem de henüz bulunabilmiş değildir. Bu nedenle biyoinformatik alanında çalışan araştırmacılar bilgisayarlı hesaplama yöntemlerinden yararlanarak bu soruna yeni çözümler bulmayı amaçlamaktadırlar. Protein yapı tahmini problemi biyoinformatiğin önemli bir çalışma alanıdır.

Bu çalışmada, hemoglobin proteininin ikincil yapı tahmini yapılmıştır. Tahmin başarısını arttırmak amacı ile Klonal Seçim Algoritmasının (KSA) sınıflandırma işlemi öncesinde uygulanması önerilmiştir. Orijinal KSA kullanımına ilave olarak algoritmada bazı değişiklikler de önerilmiştir. Sınıflandırma öncesinde iyileştirilmiş KSA ile protein verisinde iyileştirme yapılmıştır. Veri iyileştirme aşamasından sonra farklı alt yapıları kullanan farklı sınıflandırıcılar ile ayrı ayrı sınıflandırma işlemi

yapılmıştır. Sınıflandırma öncesinde orijinal KSA ve iyileştirilmiş KSA uygulanmasının tahmin başarısına etkisi incelenmiştir. Uygulama için farklı hemoglobin proteinlerini içeren bir veri seti oluşturulmuştur.

Tez çalışmasının organizasyonunda ikinci bölümde literatür araştırması yapılmış, geçmişten günümüze kadar olan süreçte protein ikincil yapı tahmini için kullanılan yöntemler tanıtılmıştır. Üçüncü bölümde proteinler, protein yapıları ve katlanmaları ile ilgili bilgi verilmiştir. Dördüncü bölümde çalışmada kullanımı önerilen KSA ve sınıflandırıcı yöntemler hakkında bilgi verilmiştir. Ayrıca uygulamadaki işlem adımları, KSA’da yapılması önerilen iyileştirmeler, KSA uygulaması için tasarlanan yazılım ve sınıflandırma için kullanılan paket program tanıtılmıştır. Beşinci bölümde çalışmada elde edilen çıktılar verilmiştir. Altıncı bölümde ise önerilen çalışmanın getirdiği avantajlar ve literatüre katkısı paylaşılmıştır.

BÖLÜM 2. LİTERATÜR ÖZETİ

Sarmal ve tabaka yapı olarak adlandırılan, aminoasit dizilerindeki düzenli katlanmaların varlığı, 1951 yılında Pauling ve ark. tarafından önerilmiştir [4, 5]. Hemoglobin ve miyoglobin proteinlerinin moleküler yapıları ilk kez Perutz ve Kendrew tarafından X-ışını kırınımı yöntemi ile çözümlenmiştir [6, 7]. Bu çalışmaları 1954 yılında Pauling'e, 1962 yılında Perutz ve Kendrew'e kimya Nobel ödülünü getirmiştir.

Rost, protein ikincil yapı tahmin yöntemlerinin tarihsel gelişimlerine göre üç nesil olarak incelenebileceğini savunmuştur [8]. 70'li yıllarda kullanılan yöntemler birinci nesil tahmin yöntemleri olarak literatürde yer alır. İkinci nesil tahmin yöntemleri 90'lı yılların başına kadar olan süreçte geliştirilmiştir. 90'lı yılların ortalarından günümüze kadar olan süreçteki tahmin yöntemleri ise üçüncü nesil yöntemler ya da teknoloji harikası yöntemler olarak adlandırılır.

Protein ikincil yapı tahmini ile ilgili, laboratuvar ortamı dışındaki ilk çalışmalar, basit istatistiksel analizlere dayanmaktadır. Az sayıdaki yapı deneysel olarak doğrulanabildiği için istatistiksel analizler de az sayıdaki gözleme göre yapılmıştır. Kısıtlı imkânların olduğu o yıllardaki tahmin başarısı %50 ile %60 arasında kalmıştır. Nagano tarafından önerilen ilk yaklaşımda, ikincil yapıların polipeptit zincirdeki altı aminoasit kalıntısına kadar olan kısa mesafeli etkileşimler ile oluştuğu var sayılmış ve kalıntı çiftlerinin ikincil yapılara karşılık gelme olasılıkları araştırılmıştır [9].

Chou ve Fasman tarafından yirmi aminoasit için protein ikincil yapılarda (sarmal, tabaka ve iplik) bulunma sıklıkları belirlenmiştir. Elde edilen bu olasılık değerlerine göre ikincil yapı tahmin kuralları oluşturulmuştur [10, 11].

İkinci nesil yöntemlerde basit istatistiksel analizlerin yerini kayan pencere yöntemi ile komşu aminoasit kalıntılarının da incelendiği koşullu olasılık çalışmaları almıştır.

Aminoasitlerin kimyasal özelliklerinin de tahmin algoritmalarında kullanılmasıyla tahmin başarısında artış sağlanmıştır fakat yine de başarı %70'lere ulaşamamıştır.

İkincil yapı tahmin çalışmalarına stereokimyasal kuralların dahil edilmesi ilk kez Lim'in çalışmalarında karşımıza çıkmaktadır. Lim, globüler proteinlerde bulunma sıklıklarına dayanarak sarmal ve tabaka yapı için geniş stereokimyasal tahmin kuralları geliştirmiştir [12, 13]. Lim'in hidrofobiklik kuralları daha doğru tahminler yapılmasına olanak sağlasa da ulaşılabilirliği daha kolay olduğu için Chou-Fasman yöntemi daha popüler olmuştur [14].

GOR yönteminde, aminoasitlerin sadece bireysel olarak ikincil yapıları oluşturma olasılıkları değil, bitişik aminoasitler ile birlikte koşullu olasılıkları dikkate alınmıştır [15]. Çalışmada on yedi pencere boyutu kullanılmıştır.

Üç boyutlu yapısı bilinen protein sayısındaki artış ve çoklu dizi hizalama yöntemleri, makine öğrenmesi algoritmalarının tahmin için kullanılması, ikincil yapı tahmin başarısının %70'i aşmasını sağlamıştır. 90'lı yıllarda yapılmış çalışmalarda tahmin başarısı %80'ler seviyesine ulaşabilmiştir.

Çoklu dizi hizalama, üç ve daha fazla sayıdaki DNA, RNA ya da protein dizilerinin karşılaştırılarak aynı sıradaki benzer bölgelerinin bulunmasıdır. Benzer bölgeler aynı kökenden gelen proteinler için korunmuş yapısal ve fonksiyonel bölgeleri ifade edebilir [16]. Bazı araştırmacılar aynı kökenden gelen proteinlerden oluşan veri setleri kullanılarak daha başarılı ikincil yapı tahmini yapılabileceğini savunmuşlardır.

Protein ikincil yapı tahmininde ilk kez çoklu dizi hizalamayı kullanan Zvelebil ve ark., homolog dizilerden oluşan bir protein ailesinden elde edilen bilgilerin kullanılmasıyla, GOR yönteminin tahmin başarısının %9 iyileştirileceğini göstermişlerdir [17].

Günümüzde kullanılan tahmin yöntemlerinin çoğunda çoklu dizi hizalama [18–23] kullanılsa da, tekli dizi tahmin yöntemleri de oldukça çalışılmış bir alandır [24–27].

Aydın ve ark. çoklu dizi hizalama kullanıldığında tahmin başarısında artış olduğunu fakat genom dizileme projelerinde tanımlanan proteinlerin önemli bir yüzdesinin bilinen her hangi bir protein ile saptanabilir bir benzerliğinin olmaması sebebi ile tekli dizi yöntemlerinin çalışılmasının daha önemli olduğunu savunmuşlardır [28, 29].

Kullanılan ikincil yapı tahmin yöntemleri ve tahmin başarısı açısından en önemli adımlardan biri 1988’de Qian ve Sejnowski’nin yapay sinir ağlarını protein ikincil yapı tahmininde uygulamaları olmuştur [30]. Bu çalışmada elde edilen başarılı sonuçların ardından makine öğrenmesi algoritmaları daha fazla tercih edilmiş ve tahmin başarılarında da artış gözlemlenmiştir [14].

Yapay sinir ağı algoritmaları, üçüncü nesil yöntemlerde ve günümüzde en fazla tercih edilen makine öğrenmesi algoritmalarından olmuştur. Araştırmacılar çalışmalarının tamamında ya da bir kısmında yapay sinir ağı algoritmalarını sıklıkla kullanmışlardır [18, 20, 31–35].

Yapay sinir ağları dışında, protein ikincil yapı tahmininde kullanılan başlıca yöntemler; saklı Markov modeli [28, 36–39], destek vektör makineleri [40, 41], arı koloni [42], karınca koloni [43], k en yakın komşu [44, 45], genetik algoritma [46, 47] olmuştur.

Bazı araştırmacılar, tahmin başarısı yüksek olan yöntemleri birlikte kullanarak ya da orijinal algoritmalarda hibritleştirme çalışmaları yaparak protein ikincil yapı tahmininde daha başarılı sonuçlar elde etmeyi hedeflemişlerdir [48–53].

Teknolojik imkânlara ulaşılabilirliğin artmasının, donanımsal maliyetlerin ucuzlamasının ve elektronik veri hacminin büyümesinin sonuçlarından biri de derin öğrenme kavramının hayatımıza girmesidir. Temeli yapay sinir ağlarına dayanan derin

öğrenme yöntemleri, 2000’li yıllardan itibaren bioinformatik alanındaki çalışmalarda sıklıkla tercih edilmiştir [54–57].

Pollastri ve ark. protein ikincil yapı tahmini için “recurrent neural networks” ve PSI-BLAST profillerini birlikte kullanarak farklı veri setleri üzerinde Q₃ ve Q₈ başarılarını incelemişlerdir [58].

Spencer ve ark., “deep learning (belief) network” yaklaşımını 3’ten 6’ya kadar gizli katman ile ayrı ayrı denemişler ve her biri 600 nörondan oluşan dört gizli katmanlı yapı ile %80,7 Q₃ başarıları elde etmişlerdir [59].

Li ve Yu, çalışmalarında deep convolutional and recurrent neural network kullanarak CB513 veri seti için %69,7 Q₈ başarıları elde etmişlerdir [60].

Yang, Heffernan ve ark., SPIDER2 adını verdikleri yöntemde, her biri yüz elli nörondan oluşan üç gizli katmanlı “deep artificial neural network” ile iteratif bir yapı kullanılırsa tahmin başarılarında %1 iyileşme olduğunu savunmuşlardır [61, 62].

Zhou ve Troyanskaya CB513 veri setinde üç katmanlı “convolutional generative stochastic network” yöntemi kullanarak %66,4 Q₈ başarıları elde etmişlerdir [63].

Wang ve ark. beş farklı veri seti üzerinde “deep convolutional neural fields” yöntemi kullanarak Q₃ ve Q₈ tahmin başarılarını hesaplamışlardır. Önerdikleri yöntem ile Q₃ başarıları %82,3 ile %85,4 arasında, Q₈ başarıları ise %68,3 ile %75,2 arasında olmuştur [64].

Wang ve ark. protein ikincil yapı tahmini için “deep recurrent encoder-decoder networks” algoritmasını önermişlerdir. Önerilen algoritma ile CB513 ve CullPDP veri setleri için Q₃ başarıları sırasıyla %82,9 ve %84,2, Q₈ başarıları %68,2 ve %73,1 olarak hesaplanmıştır [65].

Fang ve ark., protein ikincil yapı tahmini için “Deep inception-inside-inception (Deep3I) network” adını verdikleri yeni bir derin öğrenme mimarisi sunmuşlardır [66].

Protein ikincil yapı tahmini ile ilgili yapılmış olan tezler incelendiğinde, ulusal tez merkezinde kayıtlı olan tezlerin çoğunluğunun yüksek lisans tezi olduğu görülmektedir, doktora tezlerinin sayısı oldukça azdır. Bunun en büyük nedenlerinden biri büyük veri ile çalışma zorluğudur. Önerilen yöntemin literatürde kabul edilebilir olması kullanılan veri seti ile yakından ilgilidir. Veri seti büyüdükçe çalışma zorluğu artmaktadır. Büyük veri setleri kullanıldığında tahmin başarısı tek protein kullanımına göre daha düşük olmaktadır fakat veri seti kullanımı önerilen yöntemin kabul edilebilir olmasını sağlamaktadır. Yüksek lisans tezlerinin çoğunda uygulama kolaylığı nedeni ile ya tek protein için ya da küçük veri setleri için ikincil yapı tahmin çalışmaları yapıldığı görülmektedir. Doktora seviyesinde protein yapı tahmini zorlu bir süreçtir, bu nedenle doktora çalışmalarının sayısı oldukça azdır.

İbrikçi, doktora çalışmasında yapay sinir ağları ile protein ikincil yapı tahmini üzerine çalışmıştır, çalışmasında protein ikincil yapı tahmininin kolay bir alan olmadığına vurgu yapmıştır [67].

Ersöz, yirmi farklı hemoglobin proteininin A ve B zincirlerinden oluşan veri seti üzerinde, farklı yapay sinir ağı modellerinden elde ettiği sonuçları genel regresyon sinir ağları ile birleştirerek ikincil yapı tahmini yapmıştır [68]. Genel regresyon sinir ağının başarı oranı alfa zincirleri için %85,9 ile %87,3 arasında, beta zincirleri için %90,2 ile %91,7 arasında olmuştur.

Atar, farklı yapay sinir ağı algoritmaları ile hemoglobin-141, hemoglobin-146 proteinleri ve on beş proteinden oluşan veri seti kullanarak protein ikincil yapı tahmini üzerine çalışmış ve elde edilen sonuçları kıyaslamalı biçimde çalışmasında sunmuştur [69]. Hemoglobin-141 proteini için en yüksek başarı on beş pencere boyutunda SOGR (kendi kendine organize olan global sınıflamalı sıralama) algoritması kullanılarak %87,75 elde edilmiştir. Hemoglobin-146 proteini için ise en yüksek başarı on bir

pencere boyutunda GRNN (genelleştirilmiş regresyon yapay sinir ağı) kullanılarak %91,86 elde edilmiştir.

Kurt, çalışmasında CB513 veri seti kullanarak protein yapı tahmini için örüntü tabanlı kuralların bir listesini öğrenen bir algoritma geliştirmiştir [70]. Tek dizi bilgisi kullanarak %62,5, çoklu dizi hizalama kullanarak %69,2 başarı elde edilmiştir.

Demir, peptitlerin merkezi alfa karbonları arasında oluşan bağ açılarını yapay sinir ağlarının girişi olarak kullanmış ve 1bbe, 1a89, 1a9a proteinlerinin ikincil yapısını tahmin etmeye çalışmıştır [71]. Başarı oranları %49,12 ile %70 arasında olmuştur.

Arslan Tuncer, peptitlerin merkezi alfa karbonları arasında oluşan bağ açılarından faydalanarak, fidan gelişim algoritması ile 1bbe, 1a89, 1a9a proteinlerinin ikincil yapı tahmini üzerine çalışmıştır [72]. Başarı oranları %60 ile %100 arasında olmuştur.

Pakatçı, CB513 veri seti kullanarak protein ikincil yapı tahmini yapmak için saklı Markov modeli, destek vektör makinası ve doğrusal ayırtaç sınıflandırıcısı yöntemlerini birleştiren iki seviyeli bir yapı önermiştir [73]. Tahmin başarısı %67 civarında hesaplanmıştır.

Batbat, protein yapı tahmini için, aminoasitlerin hidrofobik ve hidrofilik özelliklerinden yola çıkarak oluşturulan HP modelleri ve yapay arı koloni algoritmasını kullanmıştır [74].

Akçeşme, protein dizileri arasındaki benzerliklerden faydalanarak proteinlerin ikincil yapılarını tahmin etmeye çalışmıştır [75].

Uzut, protein ikincil yapı tahmini için hibrit yöntem kullanımının tahmin başarısını arttıracığını savunmuş ve çalışmasında destek vektör makinesi, derin katlamalı yapay sinir alanları, rastgele orman yöntemlerini optimize ederek birlikte kullanmıştır [76].

Çalışmada önerilen ÇKA yöntemi tek başına ya da farklı yöntemlerle birlikte medikal alandaki sınıflandırma problemlerinde de tercih edilmiştir [77–80].

Yapay Bağışıklık konusundaki ilk çalışmalar 90’lı yılların sonunda karşımıza çıkmaktadır. Dasgupta [81], De Castro & Zuben [82] tarafından yapılan çalışmalardan bugüne kadar olan süreçte, YBS (yapay bağışıklık sistemi) bilgisayar güvenliği [83], optimizasyon [84], sürü robotiği [85], hastalık teşhisi [86] ve yolsuzluk tahmini [87] gibi farklı alanlarda uygulanmıştır.

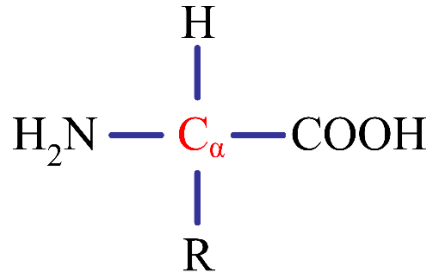
Protein yapı tahmininde ise Cutello ve ark. tarafından yapılan çalışmalar literatüre kazandırılmıştır [88–90]. Cutello ve ark. klonal seçim prensibine dayanan yapay bağışıklık algoritmalarıyla protein yapı tahmini üzerine çalışmalar yapmışlardır.

BÖLÜM 3. PROTEİNLER

Temel hücre bileşenleri karbonhidratlar, lipitler, nükleik asitler ve proteinlerdir. Proteinler hücre yapısına en fazla katılan organik moleküller olmalarının yanı sıra biyolojik faaliyetlerin tamamına yakınında da görev almaktadırlar. Enzim olarak hücredeki etkinlikleri yönetirler. Kollajen lif, kıl, keratin gibi yapılar hücreye destek oluşturur. Sinyal algılamada reseptör olarak görevlidirler. Hemoglobin kanda oksijen taşır. Plazma zarında gömülü olan birçok protein çeşitli iyon ve moleküllerin zardan geçişini sağlarlar. İnsülin kandaki glikoz seviyesini kontrol eder ve düzenler. Antikorlar bağışıklık sisteminde görevlidir. Yumurta beyazındaki albümin, tohumlardaki depo proteinler depolamada görevlidirler [91].

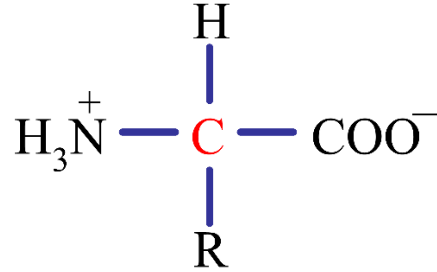
Proteinlerin yapıtaşı genetik kodun belirlediği aminoasitlerdir. Doğada üç yüzden fazla aminoasit olduğu bilinmesine rağmen bunların sadece yirmi tanesi protein yapısına katılmaktadır (bazı özel durumlarda nadiren yirmi iki olabilir). Aminoasitlerin farklı sıra ve uzunlukta bir araya gelmeleri ile farklı proteinler oluşmaktadır.

Her aminoasit α -karbon olarak adlandırılan merkezi karbon atomuna bağlı amino grubu, karboksil grubu, hidrojen atomu ve bir yan zincirden (R grubu / değişken grup) meydana gelmektedir (Şekil 3.1. ve Şekil 3.2.). Aminoasitlerin çeşitliliğini sağlayan yan zincirlerdir, farklı iki aminoasit arasındaki tek fark, yan zincirleridir [92].



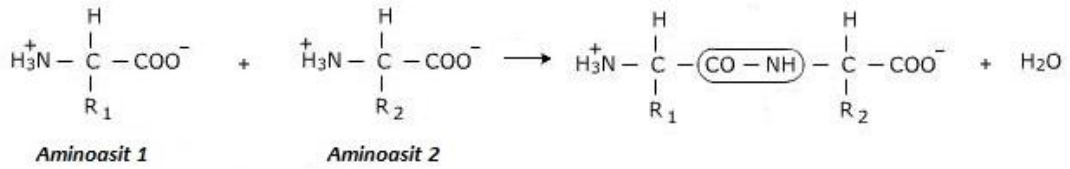
Şekil 3.1. İyonlaşmamış haldeki genel aminoasit yapısı

Fizyolojik pH'da (pH 7) amino ve karboksil grupları yüklü haldedir. Karboksil grubu bir proton (hidrojen atomu) vererek negatif yüklü olur, amino grubu ise bir proton alarak pozitif yüklü olur.



Şekil 3.2. İyonize olmuş aminoasit yapısı

Bir aminoasitin karboksil grubu ile diğer aminoasidin amino grubu, bir H₂O açığa çıkararak bir peptit bağı oluştururlar (Şekil 3.3.). İki aminoasit birleşerek dipeptit, üç aminoasit birleşerek tripeptit, çok sayıda aminoasit birleşerek polipeptit oluştururlar. Polipeptit zincirinde yer alan aminoasitler kalıntı (residue) olarak da adlandırılırlar. Üç boyutlu yapıya katlanarak bir fonksiyon gösteren polipeptitlere protein adı verilir.



Şekil 3.3. Peptit bağ oluşumu

Yan zincirler peptit bağ oluşumuna katılmazlar; proteine kutuplu, kutupsuz, hidrofilik (suyu seven), hidrofobik (sudan korkan), elektrik yüklü veya yüksüz olma gibi özellikler verirler [91]. Protein yapısına katılan yirmi aminoasit, özelliklerine göre gruplanarak Şekil 3.4.'te verilmiştir.

Yirmi aminoasidin ilk keşfedileni 1806 yılında asparajin (N), son keşfedileni 1938 yılında treonin (T) olmuştur.

Her aminoasit üç harfli bir kısaltmaya sahiptir, fakat uzun protein dizilerinde tek harfli gösterim kullanılmaktadır. Veri dosyalarının boyutunu azaltmak amacıyla biyoinformatik alanının kurucularından olan kimyager Margaret Oakley Dayhoff (1925-1983) aminoasitlerin tek harfli kodlamasını tasarlamıştır [93].

AMİNO ASİT				AMİNO ASİT			
Nonpolar, alifatik R grupları	Glisin (Gly-G) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{H} \end{array}$	Alanin (Ala-A) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH}_3 \end{array}$	Valin (Val-V) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH} \\ \\ \text{CH}_3\text{CH}_3 \end{array}$	Pozitif yüklü R grupları	Lizin (Lys-K) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH}_2 \\ \\ \text{CH}_2 \\ \\ \text{CH}_2 \\ \\ \text{NH}_3^+ \end{array}$	Arginin (Arg-R) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH}_2 \\ \\ \text{CH}_2 \\ \\ \text{CH}_2 \\ \\ \text{NH} \\ \\ \text{C}=\text{NH}_2^+ \\ \\ \text{NH}_2 \end{array}$	Histidin (His-H) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH}_2 \\ \\ \text{C}-\text{NH}^+ \\ \\ \text{C}=\text{N} \\ \\ \text{H} \end{array}$
	Lösin (Leu-L) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH}_2 \\ \\ \text{CH} \\ \\ \text{CH}_3\text{CH}_3 \end{array}$	Metiyonin (Met-M) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH}_2 \\ \\ \text{CH}_2 \\ \\ \text{S} \\ \\ \text{CH}_3 \end{array}$	İzolösin (Ile-I) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{H}-\text{C}-\text{CH}_3 \\ \\ \text{CH}_2 \\ \\ \text{CH}_3 \end{array}$		Aspartik Asit (Asp-D) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH}_2 \\ \\ \text{COO}^- \end{array}$	Glutamik Asit (Glu-E) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH}_2 \\ \\ \text{CH}_2 \\ \\ \text{COO}^- \end{array}$	
Polar, yüksüz R grupları	Serin (Ser-S) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH}_2\text{OH} \end{array}$	Treonin (Thr-T) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{H}-\text{C}-\text{OH} \\ \\ \text{CH}_3 \end{array}$	Sistein (Cys-C) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH}_2 \\ \\ \text{SH} \end{array}$	Negatif yüklü R grupları			
	Prolin (Pro-P) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_2\text{N}-\text{C}-\text{H} \\ \\ \text{H}_2\text{C}=\text{CH}_2 \end{array}$	Asparajin (Asn-N) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH}_2 \\ \\ \text{C}=\text{O} \\ \\ \text{H}_2\text{N} \end{array}$	Glutamin (Gln-Q) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH}_2 \\ \\ \text{CH}_2 \\ \\ \text{C}=\text{O} \\ \\ \text{H}_2\text{N} \end{array}$				
Nonpolar, aromatik R grupları				Nonpolar, aromatik R grupları	Fenilalanin (Phe-F) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH}_2 \\ \\ \text{C}_6\text{H}_5 \end{array}$	Tirozin (Tyr-Y) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH}_2 \\ \\ \text{C}_6\text{H}_4 \\ \\ \text{OH} \end{array}$	Triptofan (Trp-W) $\begin{array}{c} \text{COO}^- \\ \\ \text{H}_3\text{N}^+-\text{C}-\text{H} \\ \\ \text{CH}_2 \\ \\ \text{C}_8\text{H}_6\text{N} \end{array}$

Şekil 3.4. Protein yapısında bulunan 20 aminoasit

3.1. Proteinlerin Yapısı

Proteinler hiyerarşik olarak birincil, ikincil, üçüncül ve dördüncül olmak üzere dört yapıda bulunabilirler (Şekil 3.5.).

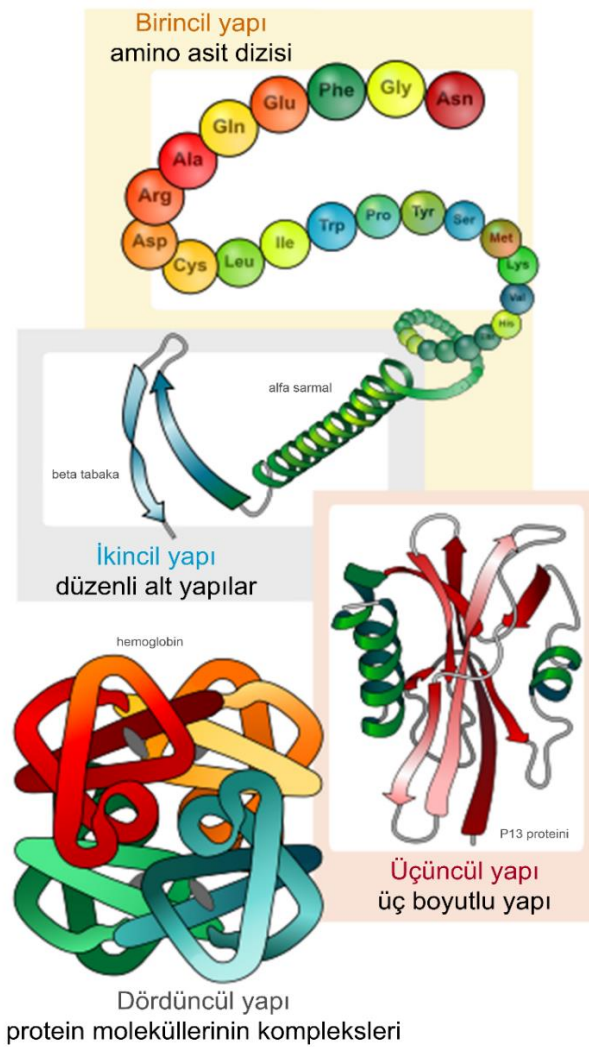
3.1.1. Birincil yapı

Aminoasitler peptit bağlarıyla bir araya gelerek proteinin birincil yapısını oluştururlar. Peptit bağlar kuvvetli bağlardır, yıkılmaları kolay değildir. Peptit bağı oluşurken daima soldaki amino asidin karboksil grubu ile sağdaki amino asidin amino grubu bağ

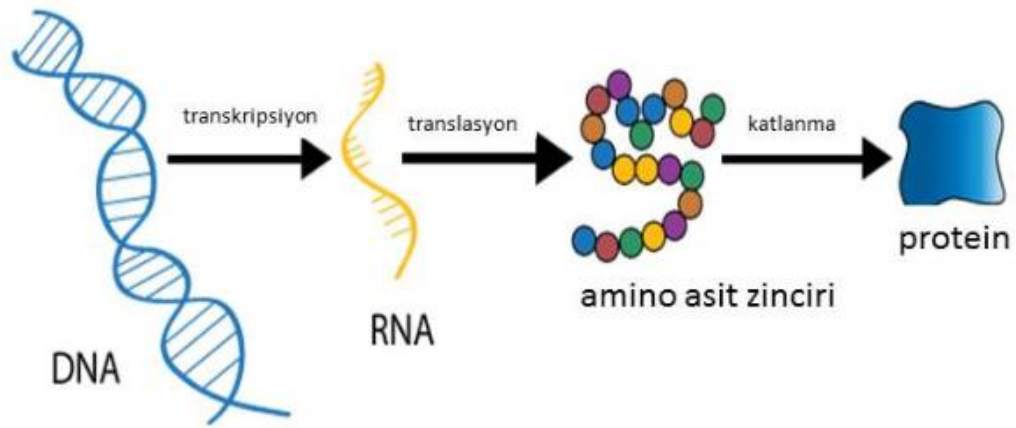
oluşturur. Polipeptit zincirinin sol ucu amino ucu ya da N-ucu, sağ ucu ise karboksil ucu ya da C-ucu olarak adlandırılır. Aminoasit dizilimi soldan sağa doğru okunur.

Hangi amino asidin hangi sırada dizileceği DNA (Deoksiribonükleik Asit) tarafından belirlenir. Tüm genetik bilgi DNA'da gizlidir, bu bilgiye göre sentezlenen RNA (Ribonükleik Asit), DNA'daki genetik şifreye göre protein sentezini yönlendirir (Şekil 3.6.).

Genetik şifrenin proteine dönüşmesi çok aşamalı ve karmaşık bir süreçtir. Bu süreçte iki önemli adım transkripsiyon (DNA'daki şifrenin RNA'ya aktarılması) ve translasyondur (şifrenin okunarak aminoasit dizisine dönüşmesi).



Şekil 3.5. Proteinlerin birincil, ikincil, üçüncül ve dördüncül yapıları [94]



Şekil 3.6. Protein sentezinde genetik bilginin akışı [95]

Nükleik asitlerin en küçük yapı birimi nükleotidlerdir. Nükleotidler, bulundurdıkları bazlara göre isim alırlar. Nükleik asitlerin yapısında beş farklı nükleotid bulunur; Adenin(A), Guanin(G) ve Sitozin(C) DNA ve RNA'da ortak bulunan bazlardır, Timin(T) sadece DNA'da, Urasil (U) ise sadece RNA'da bulunur.

Çift sarmal yapısında olan DNA zincirlerinden biri anlamlı zincir ya da kalıp zincir diğeri tamamlayıcı zincir olarak adlandırılır. İpliklerdeki A-T ve G-C bazları karşılıklı olarak hidrojen bağlarıyla bağlanarak sarmal yapıyı oluştururlar. DNA kendini eşleyebilir ve onarabilir.

RNA, DNA'daki şifreye göre sentezlenir ve bu şifreye göre protein sentezini sağlar. Üç çeşit RNA vardır; mRNA (mesajcı RNA), tRNA (taşıyıcı RNA) ve rRNA (ribozomal RNA).

mRNA'nın görevi çekirdekteki DNA'dan aldığı genetik bilgiyi, sitoplazmada bulunan ribozoma taşımaktır. mRNA sentezlenirken DNA'nın anlamlı zincirinin ilgili bölümü kalıp olarak kullanılır, DNA'daki adenin karşısına urasil, timin karşısına adenin, sitozin karşısına guanin ve guanin karşısına sitozin gelecek şekilde mRNA sentezlenir. mRNA'ya aktarılan genetik kodda her üçlü nükleotid dizisi kodon olarak adlandırılır ve her kodon bir amino asiti şifreler. Her aminoasit bir ya da daha fazla sayıda kodona karşılık gelebilir. Protein sentezi AUG kodonu ile başlar. AUG başlangıç kodonu olmasının yanı sıra, metiyonin amino asidine de karşılık gelir. UAA, UAG ve UGA ise

bitiş kodonlarıdır, bu kodonlar gelene kadar protein sentezi devam eder. Bazı özel durumlarda bitiş kodonu olarak bilinen UGA ve UAG sırasıyla 21. ve 22. aminoasitler olan Selenosistein (Sec-U) ve Prolizin (Pyl-O) sentezlenmesine neden olur. Hangi kodonun hangi aminoasite karşılık geldiği Şekil 3.7.'de verilmiştir.

İKİNCİ HARF									
Birinci Harf	U		C		A		G		Üçüncü Harf
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	"Stop"	UGA	Sec	A
	UUG		UCG		UAG	Pyl "Stop"	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met "Start"	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Şekil 3.7. Genetik (üçlü) kod aminoasit dizilimini belirler

tRNA'nın görevi aminoasitleri tanımak ve bu aminoasitleri mRNA üzerindeki kodonlara göre ribozoma taşımaktır. Kodonlara karşılık tRNA üzerinde antikodonlar bulunur.

rRNA, ribozomun yapısını oluşturur, protein sentezi sırasında peptit bağlarının kurulmasında görev alır.

Protein sentezi sırasında hangi aminoasitin hangi sırada olacağı bellidir. Protein yapısındaki tek bir aminoasitin değiştirilmesi bile proteinin yapısını ve işlevini etkiler. Örneğin mutasyon sonucu hemoglobin proteininin aminoasit diziliminde 6. sırada

glutamik asit yerine valin isimli aminoasit gelirse orak hücreli anemi denilen hastalık görülür ve hemoglobin yeterli oksijen taşıyamadığı için ölümcüldür.

Proteinlerin uzayda doğal olarak alacakları üç boyutlu yapılar (konformasyon), birincil yapılarını oluşturan aminoasitlerin dizilişlerinde gizlidir. Birincil yapıda bulunan aminoasitlerin oluşturdıkları kimyasal bağlar ve itme-çekme kuvvetleri proteinlerin üç boyutlu yapısının oluşmasında yani protein katlanmasında etkindir. Her proteinin aminoasit dizilimi ile belirlenen kendine has bir katlanma biçimi vardır. Proteinlerin üç boyutlu yapıları onların fonksiyonlarını belirler. Protein katlanmasında birincil yapı ikincil yapıyı, ikincil yapı üçüncül yapıyı, üçüncül yapı ise dördüncül yapıyı oluşturur.

3.1.2. İkincil yapı

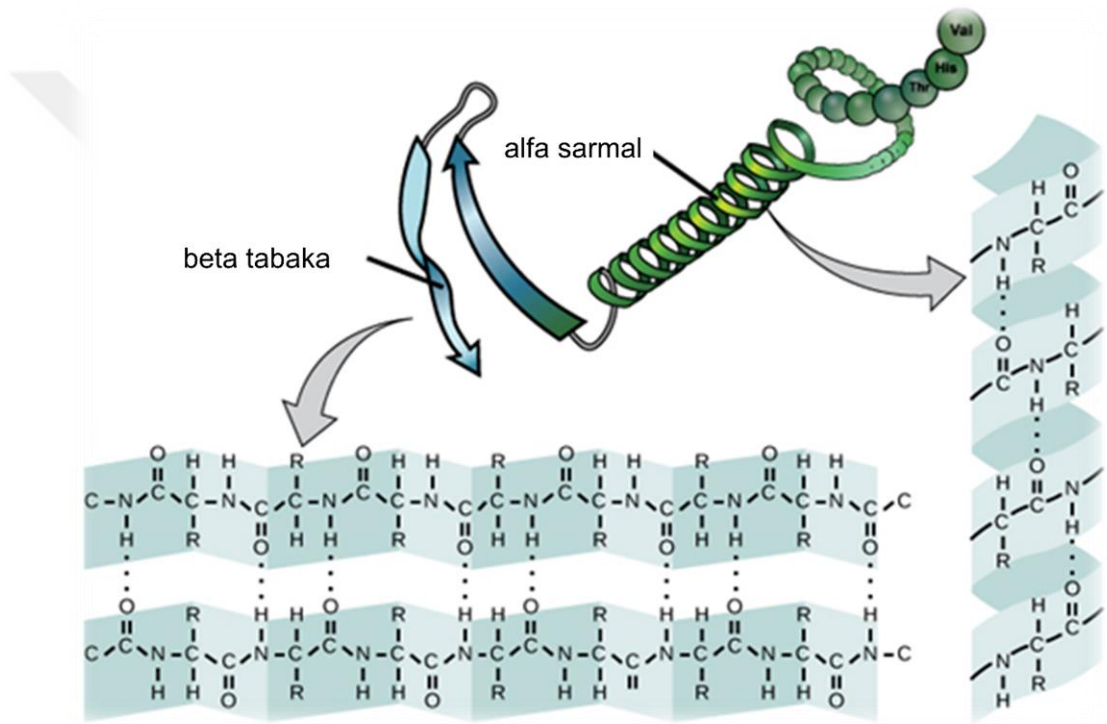
Polipeptit zincirin omurgasında bulunan amid ve karbonil grupları arasındaki hidrojen bağları ikincil yapıyı kararlı hale getirir. İkincil yapıyı oluşturan bölgesel katlanmalarda yan zincirler bağ oluşumuna katılmazlar. Temel ikincil yapılar, Şekil 3.8.'de verilmiş olan alfa sarmal (α helix) ve beta tabakadır (β sheet). Keşfedilme sırası nedeniyle sarmal yapı α , tabaka yapı ise β ile isimlendirilmiştir. Bu düzenli yapıların dışında kalan tesadüfi kıvrılmalar (random coil) ve dönüşler iplik yapı olarak adlandırılır.

- Sarmal: Proteinlerde bulunan en yaygın yapısal motiftir. Poliipeptit zincir kendi etrafında bükülerek silindir görünümlü sarmal yapıyı oluşturur. Bu bükülme, poliipeptit omurgadaki her amid grubunun, dört aminoasit uzaklıktaki karbonil grubu ile hidrojen bağı oluşturmalarıyla meydana gelir. Deri, saç ve tırnaklarda bulunan keratin sarmal şekildedir. Sarmal yapı esneklik sağlar, bu yapıdaki proteinler uzama ve kısalma yeteneğine sahiptir.
- Tabaka: Poliipeptit zincir, yan yana pileli gibi katlanmış tabaka halindedir; zincirin farklı bölümlerinin hidrojen bağlarıyla yan yana dizilmesi tabaka yapıyı oluşturur. Tabaka yapı paralel ya da anti paralel olabilir; bu yapıyı

oluşturan zincirler aynı yönde uzanıyorsa paralel, zıt yönlerde uzanıyorsa anti paralel olarak adlandırılır.

- İplik: Sarmal ve tabaka yapı dışında kalan düzensiz katlanmalar iplik yapı olarak adlandırılır.

Bir proteinin ortalama %60'ı sarmal ve tabaka yapıdan oluşur, geri kalan kısmı ise dönüşler ve düzensiz katlanmalardır [2].



Şekil 3.8. Proteinlerin ikincil yapısında bulunan motifler [96]

Çoğu proteinin ikincil yapısı hem sarmal hem de tabaka yapı içermesine rağmen bazı proteinler sadece sarmal ya da sadece tabaka yapıda bulunabilirler.

3.1.3. Üçüncül yapı

Poliipeptit zincirinin bütünsel katlamasına yani tüm aminoasitlerin üç boyutlu düzenlenmesine üçüncül yapı denir. Aminoasitlerin yan zincirlerinin etkileşimleri

katlanmaların oluşmasında etkindir. Hidrofobik aminoasitler proteinin iç bölgelerinde, hidrofilik aminoasitler ise yüzeye yakın bölgelerinde yer alırlar.

Üçüncül yapının üç ana kategorisi globüler proteinler, fibröz proteinler ve membran proteinlerdir.

3.1.4. Dördüncül yapı

Üçüncül yapının oluşumunda etkili olan bağlar ve etkileşimler dördüncül yapının oluşmasında da etkindir fakat burada üçüncül yapıdan farklı olarak, bu etkileşimlerin bir ya da daha fazla polipeptit zinciri arasında olmasıdır. Bu yapıda bulunan her protein birimine monomer ya da alt ünite adı verilmektedir. Her protein dördüncül yapıya sahip olmayabilir.

3.2. Laboratuvar Ortamında Proteinlerin Yapısal Analizi

Proteinlerin üç boyutlu yapıları hakkında bilgi edinmek için laboratuvar ortamında X-ışını kristalografisi, nükleer manyetik rezonans (NMR) spektroskopisi, kriyo (dondurucu) elektron mikroskobu ve atomik kuvvet mikroskobu yöntemleri kullanılmaktadır [1]. Laboratuvar ortamında üç boyutlu yapı analizi zorlu, maliyetli ve zaman alan bir süreç olmasının yanında, tüm proteinler için başarı ile uygulanabilen genel bir yöntem henüz yoktur. Her yöntemin kendine göre avantajları ve dezavantajları vardır.

3.2.1. X-ışını kristalografisi

1950'lerde Max Perutz ve John Kendrew'in ispermeçet balinası miyoglobininin yapısını belirlemek için kullandıkları X-ışını kristalografisi, bu çalışmayla yapısal biyolojideki yerini almıştır. Protein yapılarının belirlenmesinde kullanılan en eski tekniktir ve günümüzde hala yaygın olarak kullanılmaktadır.

Yöntemin uygulanması için öncelikle, yapısı incelenecek proteinin yeterli büyüklük, saflık ve düzendeki kristalinin elde edilmesi gerekmektedir. Her proteini kristallendirmek mümkün olmadığı için, bu yöntem tüm proteinlere uygulanamamaktadır. Belirli dalga boyuna sahip X-ışını demeti protein kristali içinden geçirilir. Protein kristalindeki atomların elektronlarına çarpan ışınlar çeşitli yönlerde dağılır. Dağılan ışınlar fotoğraf filmi ile yakalanarak kırınım modeli üretilir, film üzerinde oluşan noktaların yerleşimi ve şiddeti ölçülerek atomlar arası mesafe hesaplanır. Molekülde bulunan atomlar binlerce kırınım noktası oluşturduğu için bu hesaplamalar bilgisayar ortamında yoğun matematiksel işlemler gerektirmektedir. Bu nedenle X-ışını kristalografisi ile protein yapısının belirlenmesi zaman alan, zorlu bir süreçtir.

3.2.2. Nükleer manyetik rezonans spektroskopisi

Çözelti içindeki proteinlerin incelenebildiği bu yöntem, sadece kristallendirilebilen proteinler ile sınırlı olan X-ışını kristalografisi yöntemine göre üstünlük sağlar. Manyetik bir alana yerleştirilen yoğunlaştırılmış protein solüsyonundaki farklı atomların rezonanslarının, farklı radyo frekansları ile nasıl etkilendikleri incelenir. Herhangi bir atomun davranışı komşu atomları da etkiler, bu etkinin büyüklüğü aminoasitler arasındaki mesafenin ölçülmesini sağlar. Bu mesafeler ile proteinin üç boyutlu yapısının bir modeli oluşturulur. Hidrojenin radyo frekanslara olan yüksek duyarlılığı ve protein yapılarında bol miktarda bulunması, NMR spektrumlarının karmaşıklığını artırır. Bu karmaşıklık, büyük boyutlu proteinlerin üç boyutlu yapılarının belirlenmesini engeller.

3.2.3. Kriyoelektron mikroskopu

Kristallendirilmesi zor olan büyük proteinlerin yapısının belirlenmesi için kullanışlı bir yöntemdir. Yapısı incelenecek protein aşırı derecede düşük sıcaklıktaki (-180°C ile -260°C) sıvı helyuma daldırılıp dondurulur. Dondurma işlemi çok hızlı olduğundan su kristalleri oluşmaz, protein camsı buz içinde gömülü hale gelir ve kriyoelektron

mikroskobu ile incelenir. Görüntülerin analiz edilmesi ve proteinin üç boyutlu yapısının elde edilmesi için bilgisayar algoritmalarına ihtiyaç vardır.

3.2.4. Atomik kuvvet mikroskobu

Kantilever adı verilen çok ince uç, üç boyutlu resmi oluşturulacak yüzeye yaklaştığında bir atomik kuvvet oluşur. Kantilever yüzeye yaklaştıkça atomik kuvvet artar, uzaklaştıkça azalır. Önceden sisteme girilmiş olan kuvvet seviyesine ulaşıldığında yaklaşma işlemi durur ve kantileverin o seviyedeki pozisyonu kaydedilir. Tüm yüzey bu şekilde taranarak kantileverin pozisyonları kaydedilir ve bilgisayar ortamında üç boyutlu görüntüye dönüştürülür.

3.3. Hatalı Katlanma

Proteinlerin görevlerini doğru olarak yerine getirebilmeleri için doğru sentezlenmeleri ve üç boyutlu yapılarına doğru katlanmaları gerekmektedir. Hatalı sentezlenen ya da katlanan proteinler ölümcül genetik hastalıklara sebep olabilirler.

Alzheimer, Parkinson, tip II diyabet, katarakt, kanser, deli dana, Creutzfeldt-Jacob, kistik fibrozis, mutasyona uğrayan genetik kodun ya da hatalı katlanan proteinlerin neden olduğu hastalıkların bazılarıdır. Hatalı katlanan protein dizilerine prion adı verilir. Prionlar, etraflarındaki normal proteinleri uyarak onların da hatalı katlanmalarına ve birikmelerine neden olabilirler [97]. Prionlar merkezi sinir sistemine ait organlarda birikirse beyin fonksiyonları ve hareket sisteminde ölümcül sorunlara neden olabilirler.

Nörodejeneratif hastalıklar üzerine çalışmalar yapan Stanley Prusiner, prionları keşfiyle 1997 yılında Nobel Tıp ödülüne layık görülmüştür.

Prionlar normal proteinlere göre çok daha dirençli bir yapıya sahiptirler ve sebep oldukları hastalıkların tedavisi henüz yoktur. Bu hastalıkların bulaşıcı olma ihtimalleri tehlikenin boyutunu arttırmaktadır [98, 99]. Tüm dünyayı etkileyen ve ölümle

sonuçlanan bu hastalıkları önleyebilmek ve tedavi edebilmek için protein yapılarının bilinmesi önemli bir avantajdır. İlaç endüstrisi ve sağlık sektörünün önemli bir çalışma alanı hatalı katlanan proteinlerin neden olduğu hastalıkların tedavisidir.

Proteinlerin üç boyutlu yapıları, aminoasit dizilimleri ile belirlenir fakat bu yapıların moleküler biyologlar tarafından tahmin edilmelerinin önünde birçok zorluk vardır. Laboratuvar ortamında tüm proteinler için uygulanabilen kesin bir yöntem henüz yoktur. Protein yapı tahmini zorlu, maliyetli, zaman alan bir süreçtir ve her protein için başarılı sonuçlar elde edilememektedir. Bu nedenle biyoinformatik alanında çalışan araştırmacılar bilgisayar ortamında proteinlerin üç boyutlu yapılarını yüksek başarı oranları ile tahmin edebilmek için çalışmaktadırlar.

Çoğu proteinin üç boyutlu yapısı çok asidik ve bazik ortamlarda, yüksek sıcaklıkta ya da bazı kimyasallar ile aynı ortamda bulunduklarında değişmektedir. Buna denatürasyon denilmektedir. Laboratuvar ortamında yapılan çalışmalarda, doğal ortamlarından alınıp denatüre edilen bazı proteinlerin tekrar doğal ortamlarına konulduklarında olması gereken üç boyutlu yapılarına döndükleri gözlemlenmiştir. Bu da proteinlerin katlanmaları için gerekli olan bilginin proteinin birincil yapısında yani aminoasit diziliminde olduğunu göstermektedir.

BÖLÜM 4. MATERYAL VE YÖNTEM

Çalışmada kullanılması önerilen KSA, bağışıklık sisteminden modellenen bir algoritmadır. Bölüm 4.1.'de bağışıklık sistemi ve orijinal KSA tanıtılmıştır. Bölüm 4.2.'de bu çalışmada protein ikincil yapı tahmini için kullanılan sınıflandırıcılar hakkında bilgi verilmiştir. Bölüm 4.3.'te ise blok diyagramı Şekil 4.1.'de verilmiş olan çalışmanın ayrıntıları, çalışmada kullanılan veri, KSA'da yapılması önerilen iyileştirmeler sunulmuştur.



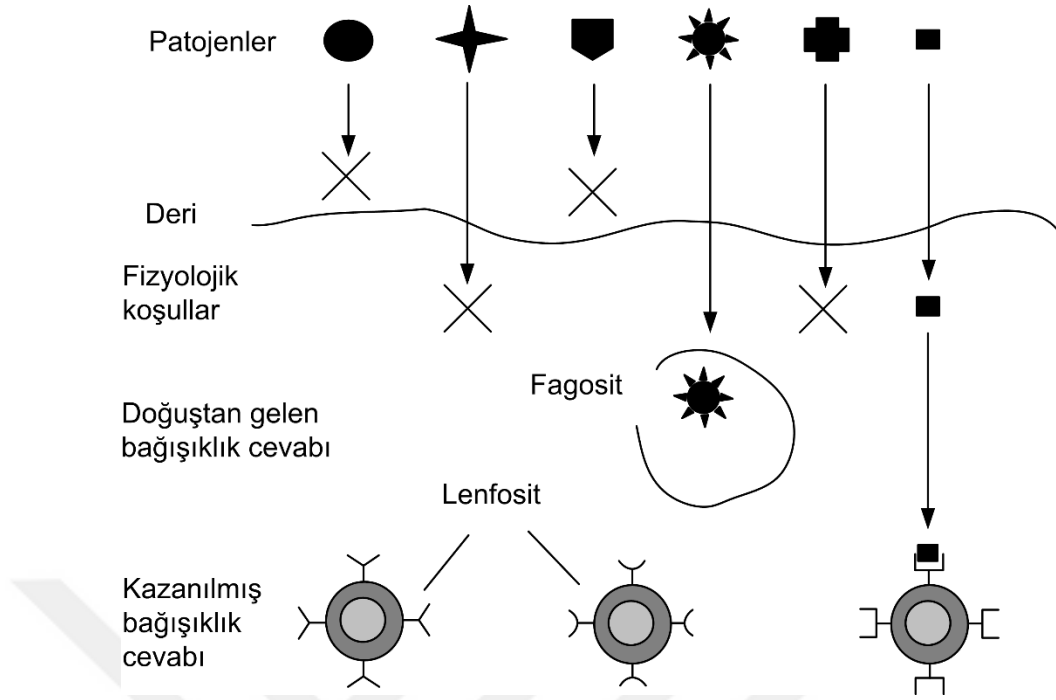
Şekil 4.1. Önerilen çalışmanın blok diyagramı

4.1. Doğal ve Yapay Bağışıklık Sistemleri

Bu bölümde doğal ve yapay bağışıklık sistemlerinin çalışma prensipleri irdelenmiş detaylı olarak nasıl çalıştıkları paylaşılmıştır.

4.1.1. Doğal Bağışıklık Sistemi

Patojen adı verilen istilacılar sürekli olarak çok hücreli canlıları tehdit ederler. Bağışıklık sistemi, bu patojenlerin vücuda girmesini engellemek ve onları yok etmek ile görevlidir. Bağışıklık sistemi Şekil 4.2.'de verilmiş olan çok katmanlı bir yapıya sahiptir. Patojenler ile ilk mücadele fiziksel ve fizyolojik bariyerler ile sağlanır; deri ve epitel yüzeyler vücut ile dış dünya arasında bir engel oluşturur. Vücut sıvıları patojenlerin vücuda tutunmalarını engeller ve içerdikleri defensin adı verilen antimikrobiyal maddeler ile patojenleri öldürürler.



Şekil 4.2. Bağışıklık sisteminin çok katmanlı yapısı [82]

Fiziksel bariyerleri aşan patojenler doğuştan gelen bağışıklık sistemi ile karşılaşırlar. Tüm patojenler için ortak bir savunma sağladığı için doğuştan gelen bağışıklık yanıtı hızlıdır fakat çok etkili değildir. Kazanılmış bağışıklık yanıtının oluşması daha uzun zaman alır, çünkü uyarıcı patojene özel savunma geliştirilir.

Bağışıklık sistemlerinin görevi konağa zarar vermeden patojenleri yok etmektir. Bunun için bağışıklık sisteminin yabancı mikroorganizmaları tanıma yeteneğinin olması gerekir.

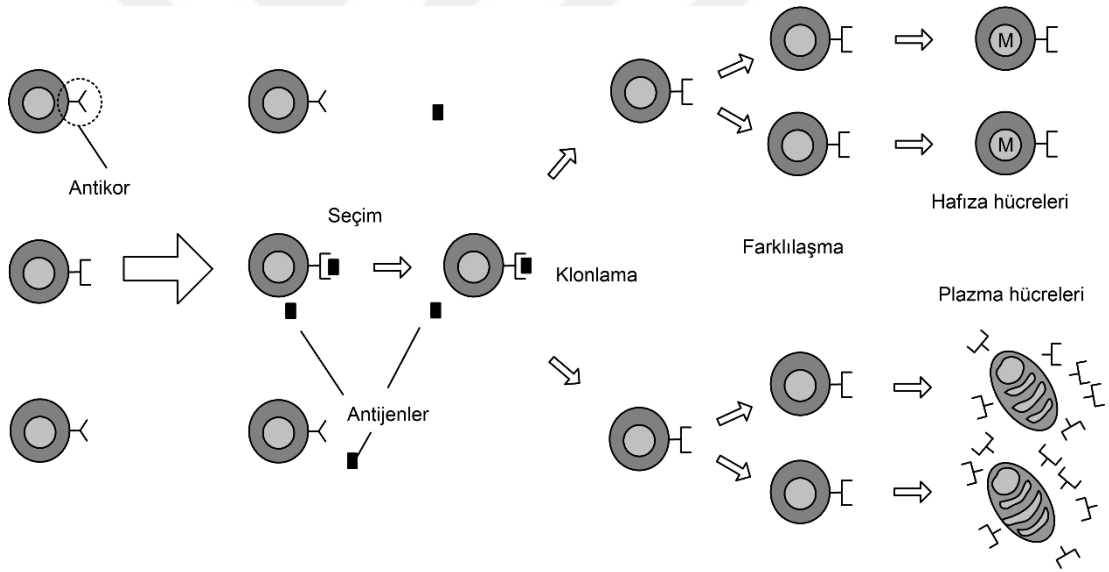
Doğuştan gelen bağışıklık, patojenleri tanıma işlemini, patojenlerde ortak olan fakat konak hücrelerde olmayan bazı özel molekülleri tanıyarak sağlar. Doğuştan gelen bağışıklık tek bir patojene özel değil, tüm patojenlere karşı genel bir savunmadır, bu nedenle bağışıklık yanıtı çok hızlı gelişir; patojenleri yok etmek için hemen etkinleşen proteinler ve fagositik hücreler üretilir.

Doğuştan gelen bağışıklık yanıtı, kazanılmış bağışıklık yanıtını da etkinleştirir. Doğuştan gelen bağışıklık ve kazanılmış bağışıklık birbiri ile etkileşim halindedir.

Kazanılmış bağışıklık, belirli bir patojene özgüdür ve doğal bağışıklığa göre çok daha güçlü bir savunma sağlar. Kazanılmış bağışıklık yanıtını tetikleyen maddelere antijen adı verilir. Antijenlerin özelliklerine göre antikor üretilir ve bu antikorlar antijenleri yok etmek için çalışırlar.

Kazanılmış bağışıklık yanıtı, lenfosit adı verilen akyuvarlar tarafından sağlanır. Lenfositler B hücreleri ve T hücreleri olarak iki farklı gruba ayrılır. B lenfositleri antikor üretiminden sorumludurlar, kemik iliğinde olgunlaşırlar. T lenfositlerinin farklı alt türleri vardır ve bunlar hücresel savunma ve bağışıklıkta görevlidir.

Kazanılmış bağışıklık sisteminin farklı patojenlere özel savunma sağlaması klonal seçim prensibine dayanmaktadır (Şekil 4.3.).



Şekil 4.3. Klonal seçim prensibi

Organizmada bir antijen tespit edildiğinde, o antijeni tanıyan B hücreleri uyarılır. Uyarılan B hücreleri çoğalırlar ve antijenleri tanıma kabiliyetlerini arttırmak amacı ile farklılaşma sürecine girerler. B lenfositlerinin yüzeylerindeki antikorlar sadece bir çeşit antijene bağlanarak onları tanıyabilirler. Aynı antijenle daha sonra tekrar karşılaşma ihtimaline karşı bazı lenfositler hafıza hücresi olarak bağışıklık sisteminde saklanırlar. Lenfosit repertuarı denilen bu yapı sayesinde, kazanılmış bağışıklık sistemi hafızaya sahip olur, aynı antijen ile ikinci kez karşılaşıldığında ilkinde göre daha kısa

sürede bağışıklık cevabı oluşur. Güçlü bir savunma için lenfosit repertuarı sürekli güncellenmektedir. Lenfosit repertuarı ne kadar geniş ise bağışıklık o kadar kuvvetlidir. İnsanların bağışıklık sisteminde yaklaşık 10^{12} adet B ve T lenfosit vardır.

4.1.2. Yapay Bağışıklık Sistemi

Her bireyin kendine has, özelleşmiş bir bağışıklık sistemi vardır. Bağışıklık sisteminin merkezi bir kontrol mekanizması yoktur, tüm vücut boyunca dağıtılmış bir yapıdadır. Zararlı ve zararsız molekülleri ayırt edebilir. İlk kez karşılaştığı patojenlere karşı savunma geliştirebilir, bu patojenlerin yapısını öğrenip hafızasında tutabilir ve bu sayede aynı patojen ile tekrar karşılaştığında daha hızlı ve güçlü yanıt oluşturabilir.

Bağışıklık sisteminin bu önemli özellikleri bilgisayar bilimcilerinin dikkatini çekmiştir. Biyolojik sistemlerden ve doğadan ilham alan diğer yöntemlerin problem çözmede sağladığı avantajlar ve başarılar, bağışıklık sisteminin de modellenmesi düşüncesini doğurmuştur. Bağışıklık sisteminin her bir işlevsel yapısının modeli, yapay bağışıklık sistemi (YBS) adı altında incelenmektedir. Klonal seçim algoritması (KSA), negatif seçim algoritması, bağışıklık ağı teorisi YBS çatısı altında yer alan en temel tekniklerdir [100].

4.1.2.1. Klonal seçim algoritması

KSA, biyolojik klonal seçim prensibinden esinlenerek modellenmiştir. Klonal seçim prensibinde amaç, antijenlerle mücadele edebilecek antikor çeşitliliğini sağlamaktır. Her yeni antijen ile karşılaşıldığında, bağışıklık ağı bu antijenlere göre güncellenerek antijenlerin tanınırlığı artırılır.

Antijenler (Ag), çözülmesi beklenen problem kümesinin; antikorlar (Ab) ise çözüm kümesinin elemanlarıdır. Protein ikincil yapı tahmini probleminde, ab ve ag protein yapısında bulunan yirmi çeşit aminoasit olmak üzere, Ab ve Ag, L uzunluklu vektörler ile temsil edilmektedirler. Antikor popülasyonu (P), n adet antikordan oluşur.

Ag, Ab ve P kümeleri, sırası ile Denklem 4.1, 4.2 ve 4.3'te paylaşılmıştır.

$$Ag = \{ ag_1, ag_2, ag_3, \dots, ag_L \} \quad (4.1)$$

$$Ab = \{ ab_1, ab_2, ab_3, \dots, ab_L \} \quad (4.2)$$

$$P = \{ Ab_1, Ab_2, Ab_3, \dots, Ab_n \} \quad (4.3)$$

Bir antijenin tanınabilmesi için, antijen ve antikor bağlanma bölgelerinin birbirine uygun olması gerekmektedir. Benzerlik denilen bu uygunluk değeri, antijen ve antikor arasındaki uzaklık ile ters orantılıdır. Antijen ve antikorlar sayısal değerlerden oluşan vektörler ile temsil ediliyorsa Öklid ya da Manhattan gibi uzaklık ölçütleriyle, semboller ya da ikili ifadelerden oluşan vektörlerle temsil ediliyorsa Hamming uzaklık ölçütü ile benzerlik hesabı yapılabilir.

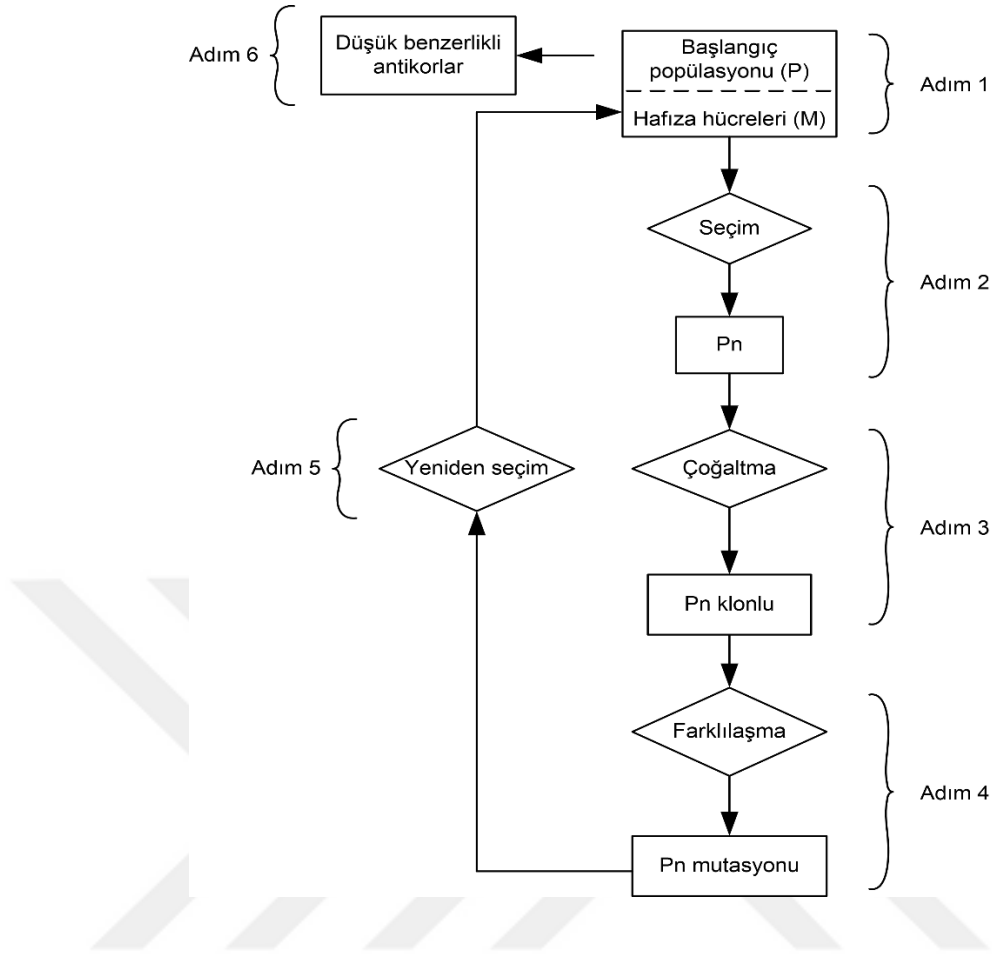
Sözü geçen Öklid, Manhattan ve Hamming uzaklıkları sırası ile Denklem 4.4, 4.5 ve 4.6'da verilen eşitlikler ile hesaplanır.

$$\text{Öklid } (Ag, Ab) = \sqrt{\sum_{i=1}^L (ab_i - ag_i)^2} \quad (4.4)$$

$$\text{Manhattan } (Ag, Ab) = \sum_{i=1}^L |ab_i - ag_i| \quad (4.5)$$

$$\text{Hamming } (Ag, Ab) = \sum_{i=1}^L \delta_i, \quad \delta_i = \begin{cases} 1, & ab_i \neq ag_i \text{ ise} \\ 0, & ab_i = ag_i \text{ ise} \end{cases} \quad (4.6)$$

Yukarıda ayrıntıları verilmiş olan KSA'nın işlem adımları, Şekil 4.4.'teki blok diyagramda sunulmuştur.



Şekil 4.4. Klonal seçim algoritması blok diyagramı

KSA'nın işlem adımları şu şekilde verilebilir [101];

- Adım 1: Veri setinden antikor popülasyonunu oluştur.
- Adım 2: Hamming uzaklık ölçütü kullanarak her bir antijene en yakın antikoru seç.
- Adım 3: Seçilen antikorları çoğalt.
- Adım 4: Mutasyon oranı kadar sütunda mutasyon yap, mutasyonlu antikorları aday antikorlar kümesine kaydet.
- Adım 5: Aday antikorlar ile antijenlerin benzerliklerini hesapla. Eşik değerin üzerindeki benzerliğe sahip antikorları hafıza hücresi olarak başlangıç popülasyonuna dahil et.

Adım 6: Düşük benzerliğe sahip olan antikorları başlangıç popülasyonundan çıkart.

KSA'nın işlem adımlarını daha iyi anlamak ve buna bağlı olarak sonuçlarını görmek açısından Özkan [104] tarafından kullanılmış olan gözlem tablosu incelenebilir (Tablo 4.1.). Gözlem tablosundaki X1, X2 nitelikleri ve Y sınıfından oluşan örnek veri için KSA işleyişi görülebilir.

Tablo 4.1. KSA için örnek gözlem tablosu

Gözlem No	X1	X2	Y
1	2	4	KÖTÜ
2	3	6	İYİ
3	3	4	İYİ
4	4	10	KÖTÜ
5	5	8	KÖTÜ
6	6	3	İYİ
7	7	9	İYİ
8	9	7	KÖTÜ
9	11	7	KÖTÜ
10	10	2	KÖTÜ

Gözlem tablosundaki verilerin %70'i eğitim, %30'u test verisi olarak rasgele ayrılmıştır. Gözlem no 7, 9 ve 10 olan satırlar test verisi olarak seçilmiştir. Eğitim ve test olarak ayrılan veriler Tablo 4.2. ve Tablo 4.3.'te verilmiştir.

Tablo 4.2. Eğitim için ayrılan veri (Antikorlar)

X1	X2	Y
2	4	KÖTÜ
3	6	İYİ
3	4	İYİ
4	10	KÖTÜ
5	8	KÖTÜ
6	3	İYİ
9	7	KÖTÜ

Tablo 4.3. Test için ayrılan veri (Antijenler)

X1	X2	Y
7	9	İYİ
11	7	KÖTÜ
10	2	KÖTÜ

Veriler sayısal olduğu için benzerlik hesabı Öklid uzaklık ölçütü kullanarak yapılabilir. Test setindeki her bir verinin eğitim setindeki tüm verilere uzaklıkları ayrı ayrı hesaplanır. Aşağıda üç antijen için benzerlik hesabı verilmiştir.

(7, 9) Antijeni ile Antikorlar arasındaki uzaklıklar:

- Öklid $((7,9), (2,4)) = \sqrt{(2-7)^2 + (4-9)^2} = 7,07$
- Öklid $((7,9), (3,6)) = \sqrt{(3-7)^2 + (6-9)^2} = 5$
- Öklid $((7,9), (3,4)) = \sqrt{(3-7)^2 + (4-9)^2} = 6,4$
- Öklid $((7,9), (4,10)) = \sqrt{(4-7)^2 + (10-9)^2} = 3,16$
- Öklid $((7,9), (5,8)) = \sqrt{(5-7)^2 + (8-9)^2} = 2,24$
- Öklid $((7,9), (6,3)) = \sqrt{(6-7)^2 + (3-9)^2} = 6,08$
- Öklid $((7,9), (9,7)) = \sqrt{(9-7)^2 + (7-9)^2} = 2,83$

(11, 7) Antijeni ile Antikorlar arasındaki uzaklıklar:

- Öklid $((11,7), (2,4)) = \sqrt{(2-11)^2 + (4-7)^2} = 9,49$
- Öklid $((11,7), (3,6)) = \sqrt{(3-11)^2 + (6-7)^2} = 8,06$
- Öklid $((11,7), (3,4)) = \sqrt{(3-11)^2 + (4-7)^2} = 8,54$
- Öklid $((11,7), (4,10)) = \sqrt{(4-11)^2 + (10-7)^2} = 7,61$
- Öklid $((11,7), (5,8)) = \sqrt{(5-11)^2 + (8-7)^2} = 6,08$
- Öklid $((11,7), (6,3)) = \sqrt{(6-11)^2 + (3-7)^2} = 6,40$
- Öklid $((11,7), (9,7)) = \sqrt{(9-11)^2 + (7-7)^2} = 2$

(10, 2) Antijeni ile Antikorlar arasındaki uzaklıklar:

- Öklid $((10,2), (2,4)) = \sqrt{(2-10)^2 + (4-2)^2} = 8,25$
- Öklid $((10,2), (3,6)) = \sqrt{(3-10)^2 + (6-2)^2} = 8,06$
- Öklid $((10,2), (3,4)) = \sqrt{(3-10)^2 + (4-2)^2} = 7,28$
- Öklid $((10,2), (4,10)) = \sqrt{(4-10)^2 + (10-2)^2} = 10$
- Öklid $((10,2), (5,8)) = \sqrt{(5-10)^2 + (8-2)^2} = 7,81$
- Öklid $((10,2), (6,3)) = \sqrt{(6-10)^2 + (3-2)^2} = 4,12$
- Öklid $((10,2), (9,7)) = \sqrt{(9-10)^2 + (7-2)^2} = 5,10$

Antijenler ve antikorlar arasındaki uzaklık ne kadar küçük ise benzerlik o kadar yüksektir. (7,9), (11,7) ve (10,2) antijenlerine en küçük mesafedeki antijenler sırasıyla (5,8),(9,7) ve (6,3)'tür. Yani bu antijenler en yüksek benzerliğe sahiptirler ve Tablo 4.4.'te verilmiş olan P_n kümesini oluştururlar. Seçilen bu en iyi bireyler antijene olan benzerliklerine bağlı olarak çoğalır. Benzerliği yüksek olan bireyler daha fazla sayıda çoğalacaktır, benzerliği düşük olan bireyler daha az sayıda çoğalacaktır. Bu örnekte tüm bireyler eşit sayıda çoğaltılmıştır (Tablo 4.5.).

Tablo 4.4. P_n tablosu

X1	X2	Y
5	8	KÖTÜ
9	7	KÖTÜ
6	3	İYİ

Tablo 4.5. P_n klonlu tablosu

X1	X2	Y
5	8	KÖTÜ
5	8	KÖTÜ
9	7	KÖTÜ
9	7	KÖTÜ
6	3	İYİ
6	3	İYİ

Mutasyon oranı, kaç sütunun mutasyona uğrayacağını belirten değerdir. Burada mutasyon oranı 1 olarak alınmış yani 1 sütun mutasyona uğratılmıştır (Tablo 4.6.).

Tablo 4.6. Pn mutasyonlu tablosunun oluşturulması

X1	X2	Y
5	7	KÖTÜ
5	3	KÖTÜ
9	8	KÖTÜ
9	3	KÖTÜ
6	8	İYİ
6	7	İYİ

Çoğaltma ve farklılaşmadan sonra gelişmiş bireyleri popülasyona eklemek için yeniden seçme işlemi yapılır ve M hafıza seti oluşturulur. Antijenlerin Pn mutasyonlu ile benzerlikleri hesaplanır, eşik değerinin üzerindeki benzerliğe sahip olanlar hafıza hücresi olarak başlangıç popülasyonuna katılırlar. Başlangıç popülasyonundaki düşük benzerlikteki bireyler popülasyondan çıkartılabilir.

Pn mutasyonlu kümesinde, (7,9) antijenine en yakın olan (6,8), (11,7) antijenine en yakın olan (9,8), (10,2) antijenine en yakın olan (9,3) olduğu için bu veriler P popülasyonuna dahil edilir. P popülasyonundaki (2,4) antikoru üç antijene de en uzak antikoru olduğu için P popülasyonundan çıkartılır (Tablo 4.7.).

Tablo 4.7. KSA ile eğitilmiş veri

X1	X2	Y
2	4	KÖTÜ
3	6	İYİ
3	4	İYİ
4	10	KÖTÜ
5	8	KÖTÜ
6	3	İYİ
9	7	KÖTÜ
6	8	İYİ
9	8	KÖTÜ
9	3	KÖTÜ

Yapılan bu işlemler sonucunda veriler KSA ile eğitilmiştir. Çoğaltma ve farklılaşma işlemleri antikorların antijenlere olan duyarlılıklarını arttırmış olup (daha fazla sayıda ve daha sağlıklı veriler elde edilmesi ile) sınıflandırma başarısında artış olacaktır.

KSA'nın yukarıda verilen işlem adımları ile lenfosit repertuarını devamlı olarak güncel tutması ve antijen tanıma kapasitesini arttırması sayesinde, sınıflandırma problemlerinin çözümündeki başarı oranlarını yukarıya çekebileceği düşünülmüştür. Bu nedenle bu tez çalışmasında proteinlerin ikincil yapılarının tahmini öncesinde, KSA uygulanarak tahmin başarısının arttırılması hedeflenmiştir.

4.1.2.2. Negatif seçim algoritması

Bağışıklık sisteminin yabancı molekülleri ve kendine ait molekülleri ayırt edebilmesi önemli bir konudur. Bu ayırım sağlıklı bir şekilde yapılamazsa otoimmün hastalıklar ile karşı karşıya kalınır.

T hücrelerinin olgunlaşması Timüs bezinde gerçekleşir. Bu süreçte, kendi hücrelerini tanıyan reseptörlere sahip olan T hücreleri tespit edilirse bunlar imha edilir. Kendi hücrelerini tanımayan reseptörlere sahip T hücreleri ise Timüsten tüm vücuda yayılır. Negatif seçim prensibi olarak adlandırılan bu süreçte amaç bağışıklık sisteminin kendi hücrelerine saldırmasını engellemektir. Hayatta kalan T hücreleri sadece yabancı hücreleri tanıyacak ve onlarla reaksiyona girecektir.

Negatif seçim algoritması, biyolojik negatif seçim prensibinin matematiksel formülasyonudur. Bölüm 4.1.2.1.'de verilmiş olan antijen-antikor gösterim şekilleri ve uzaklık ölçütleri bu algorithmada da geçerlidir.

Algoritmanın işlem adımları şu şekildedir:

- Öz hücre kümesini belirle.
- Öz hücre kümesine ait olmayan elemanları tanıyacak aday detektörler kümesi oluştur.
- Aday detektörler ile öz hücre kümesinin elemanlarını karşılaştır.

- Öz hücrelerle benzerliği eşik değerinin üzerinde olan detektörleri yok et.
- Öz hücrelerle eşleşmeyen detektörleri yeni karşılaşılan hücrelerin anomali durumunu tespit etmek için kullan.
- Yeni hücreler ile detektörler arasında eşleşme varsa anormal durum, eşleşme yoksa normal durum olarak belirle.

Negatif seçim algoritması bilgisayar güvenliği, anomali ve hata tespiti problemleri için oldukça uygun bir yaklaşımdır.

4.1.2.3. Bağışıklık ağı teorisi

Antijenler üzerinde bulunan tanımlayıcı bölgeler epitop olarak adlandırılır. Antikorların antijenleri tanıyabilmesi için paratop bölgeleri ile antijenlerin epitop bölgelerinin uyumlu olması gerekmektedir. Antikorlarda bulunan epitoplar ise idiotop olarak adlandırılır. İdiotoplar antikorların birbirini algılamasını ve birbirleriyle haberleşmesini sağlar. Antijene duyarlı lenfositler tanıma sinyaline pozitif veya negatif olarak cevap verebilir. Olumlu bir tepki hücre çoğalmasına, hücre aktivasyonuna ve antikor salgılanmasına yol açarken, olumsuz bir tepki tolerans ve baskılanmaya yol açacaktır.

Bağışıklık sistemi, antijenlerin yokluğunda bile birbirini tanıyan düzenli moleküller ve hücreler ağından oluşmaktadır. Bu şekilde hücreler arası iletişim ve öğrenme gerçekleşir.

4.2. Sınıflandırıcılar

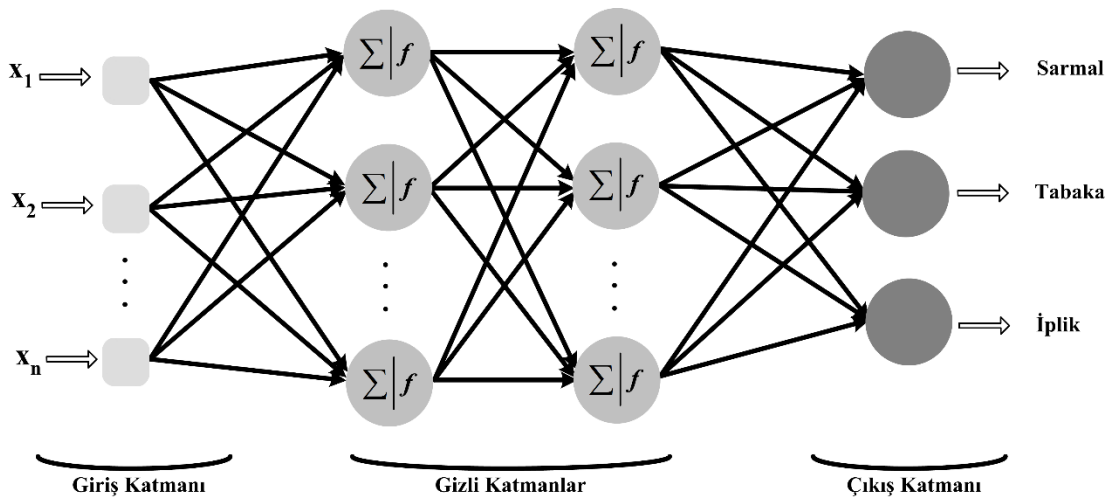
Farklı tekniklere dayanan sınıflandırma algoritmaları ile ayrı ayrı sınıflandırma yapılmıştır. Mesafeye dayalı sınıflandırma algoritması kNN, istatistiksel sınıflandırma algoritmalarından Naive Bayes, Karar ağaçları temelli ID3 ve Rapid Miner'ın derin öğrenme başlığı altında bulunan Çok Katmanlı Algılayıcılar (ÇKA) sınıflandırma için seçilen yöntemlerdir.

4.2.1. Çok Katmanlı Algılayıcı

Çok katmanlı algılayıcı ağlar (ÇKA), yapay sinir ağlarının öğretmenli öğrenme (supervised) stratejisini kullanırlar. Öğretmenli öğrenmede, her örnek için girdiler ve o girdilerden elde edilmesi gereken çıktılar sisteme verilir. Sistem girdiler ile çıktılar arasındaki ilişkileri öğrenir ve daha sonra benzer örnekler için çözüm üretebilir [102].

Rumelhart ve ark. tarafından geliştirilen ÇKA modeline hata yayma modeli ya da geriye yayılım modeli de denmektedir [103]. Ağın beklenen çıktısı ile ürettiği çıktı arasındaki hatayı en aza indirmek için, hata ağa yayılır.

ÇKA yapısında üç farklı katman yer almaktadır; giriş katmanı, ara katmanlar (gizli katmanlar) ve çıkış katmanı. Giriş katmanında bilgi işleme olmaz, gelen veriler ara katmana iletilir. Bu katmandaki her nöron bir sonraki katmanda yer alan tüm nöronlara bağlıdır. Giriş katmanından alınan veriler ara katmanda işlenir. Katman sayısı öğrenme katsayısını temsil eder, fakat katman sayısı ne kadar fazla olursa öğrenme o kadar iyi olur gibi bir genelleme yapılamaz. Çıkış katmanında, ara katmandan alınan bilgilere göre çıkış değerleri elde edilir. İşlemler ağda, giriş katmanından çıkış katmanına doğru adım adım yürütülür. ÇKA ileri beslemeli sinir ağıdır. ÇKA yapısı Şekil 4.5.'te gösterilmektedir.



Şekil 4.5. ÇKA yapısı

YSA yapısında bilgi nöronlarda tutulur, nöron sayısı belleği temsil eder. Her nöronun çıktısı Denklem 4.7.'de verilmiş olan eşitlik kullanılarak hesaplanır.

$$f(\alpha) = f(\sum_{i=1} w_i x_i + b) \quad (4.7)$$

Denklemde x_i ve w_i sırasıyla nöronun giriş değerlerini ve onların ağırlıklarını temsil eder. f ağ boyunca kullanılan doğrusal olmayan aktivasyon fonksiyonudur. Bias değeri b , nöronun aktivasyon eşiğini temsil eder. Çok katmanlı yapay sinir ağlarında amaç, kurulan model için en iyi sonucu üretecek ağırlık ve bias değerlerinin bulunmasıdır. Her epochta ağırlıklar ve bias güncellenir. Bu değerlerin hesaplanma süreci öğrenme olarak adlandırılır. Aktivasyon fonksiyonu olarak tanh, rectifier ve maxout tercih edilebilir. Bu fonksiyonlar sırası ile Denklem 4.8, 4.9 ve 4.10'da verilen eşitlikler kullanılarak hesaplanır.

$$\tanh(\alpha) = \frac{e^{\alpha} - e^{-\alpha}}{e^{\alpha} + e^{-\alpha}} \quad (4.8)$$

$$\text{rectifier}(\alpha) = \max(0, \alpha) \quad (4.9)$$

$$\text{maxout}(\alpha_1, \alpha_2) = \max(\alpha_1, \alpha_2) \quad (4.10)$$

Hatayı minimize etmek için ağırlıklar ve bias her epochta Denklem 4.11 ve 4.12 kullanılarak güncellenir. Burada α öğrenme oranı adı verilen sabit bir parametredir.

$$w_{jk} := w_{jk} - \alpha \frac{\partial L(W, B | j)}{\partial w_{jk}} \quad (4.11)$$

$$b_{jk} := b_{jk} - \alpha \frac{\partial L(W, B | j)}{\partial b_{jk}} \quad (4.12)$$

Karmaşıklığı artan bir problemin yapay sinir ağları ile çözülmesi katman sayısının artırılması ile mümkündür. Artan katman sayısı, daha derin bir karar mekanizması anlamına gelmekle birlikte daha fazla işlemci gücü de demektir. Derin öğrenme son

yıllarda oldukça popüler olmasına rağmen aslında yeni bir konu değildir. Çağımızdaki teknolojik gelişmeler derin öğrenme kavramını literatüre sokmuştur.

4.2.2. K-En Yakın Komşu Algoritması

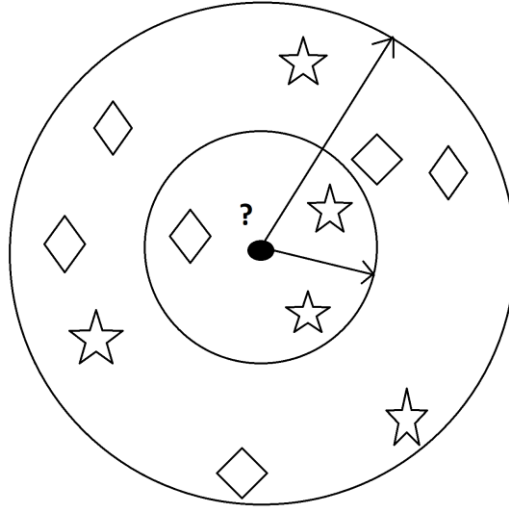
Mesafeye dayalı sınıflandırma algoritmalarından olan k-en yakın komşu algoritması (kNN) uygulama kolaylığı nedeni ile tercih edilen, yaygın kullanıma sahip bir algoritmadır.

Sınıfları belli olan bir veri kümesindeki gözlem değerlerinden yararlanılarak, sınıfı bilinmeyen yeni bir gözlem değerinin ait olduğu sınıf kNN yöntemi ile belirlenebilir [104].

kNN ile sınıflandırma yapılırken işlem adımları şu şekildedir:

- Öncelikle k değeri belirlenir, k değeri en yakın komşu sayısını ifade eder.
- Sınıfı belirlenecek olan gözlem değeri ile sınıfı önceden bilinen tüm değerlerin mesafeleri hesaplanır.
- En küçük mesafeye sahip k adet veri belirlenir ve bunların çoğunluğunun dahil olduğu sınıf seçilir.
- Seçilen sınıf, gözlem değerinin ait olduğu sınıf olarak kabul edilir.

kNN algoritmasında k değerinin belirlenmesi önemlidir; k çok büyük seçilirse özellikleri çok benzemeyen elemanlar aynı kümeye dahil edilebilir. Şekil 4.6.'da seçilen farklı k değerleri için farklı sınıflara atama yapılabileceği görülmektedir.



Şekil 4.6. kNN algoritmasında k değerinin belirlenmesi

Sınıflandırma algoritması olarak KSA ile birlikte genellikle kNN tercih edilmektedir.

4.2.3. Naive Bayes Sınıflandırıcısı

Bayes sınıflandırıcıları istatistiksel sınıflandırma teknikleri arasındadır, Denklem 4.13.'te verilen Bayes Teoremi bu sınıflandırıcının temelini oluşturur. Koşullu olasılıklardan yararlanarak sınıfı bilinmeyen gözlem değerinin sınıfı belirlenir.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (4.13)$$

Bu denklemde $P(H|X)$, X olayı gerçekleştiği durumda H olayının gerçekleşme olasılığıdır, $P(X|H)$, H olayı gerçekleştiği durumda X olayının gerçekleşme olasılığıdır, $P(H)$ ve $P(X)$, H ve X olaylarının gerçekleşme olasılığıdır. $P(H)$ ve $P(X)$ önsel olasılık, $P(X|H)$ sonrasal olasılık olarak tanımlanır.

X ait olduğu sınıf bilinmeyen ve $X=\{x_1, x_2, \dots, x_n\}$ nitelik değerlerinden oluşan örnek veri olsun. $C_1, C_2, \dots, C_n$ sınıf değerleri olsun. Sınıfı belirlenecek olan örnek için Denklem 4.14 değeri hesaplanır.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (4.14)$$

$P(X|C_i)$ olasılık hesabı için x_i değerlerinin birbirinden bağımsız olduğu kabul edilirse işlemler daha kolay yapılabilir ve Denklem 4.15 kullanılabilir.

$$P(X|C_i) = \prod_{k=1}^n P(X_k|C_i) \quad (4.15)$$

Denklem 4.14'teki $P(X)$ değeri tüm $P(C_i|X)$ 'ler için eşit olduğundan sadece $P(X|C_i)P(C_i)$ değerlerini karşılaştırmak yeterlidir. Bu değerlerden en büyük olanı X 'in sınıfı olacaktır.

$$\arg \max_{C_i} \{P(X|C_i)P(C_i)\} \quad (4.16)$$

Sonrasal olasılıkları kullanan Denklem 4.16 en büyük sonrasal sınıflandırma yöntemi (Maximum A Posteriori classification=MAP) olarak da bilinir. Bayes sınıflandırıcısı olarak Denklem 4.17 kullanılabilir [104].

$$C_{MAP} = \arg \max_c \prod_{k=1}^n P(X_k|C_i) \quad (4.17)$$

4.2.4. ID3 Algoritması

Quinlan tarafından geliştirilen ID3 (Iterative Dichotomiser 3) algoritması ile sınıflandırma karar ağaçları yardımıyla yapılır [105]. Oluşturulan karar ağacı sınıfı bilinmeyen örneklerin sınıflarını belirlemek için kullanılır. Karar ağacında dallanmanın hangi niteliğe göre yapılacağı entropi ya da bilgi kazanımı ile belirlenir. Tüm özniteliklerin entropisi ya da bilgi kazanımı hesaplanır. En küçük entropi ya da en büyük bilgi kazanımına sahip olan öznitelik ile bölünmeye başlanır. Kalan her öznitelik için aynı işlemler tekrarlanır ve diğer dallanmalar belirlenir.

Entropi, veri kümesindeki belirsizliği ölçmek için kullanılır ve 0-1 arasında bir değer alır. $P = \{p_1, p_2, \dots, p_n\}$ olasılıklar olmak üzere tüm bu olasılıkların toplamı 1 olmalıdır. Bu durumda entropi Denklem 4.18'deki gibi hesaplanacaktır.

$$H(p_1, p_2, \dots, p_n) = -\sum_{i=1}^n p_i \log_2(p_i) \quad (4.18)$$

Veri bölünmeden önce doğru sınıflandırma yapmak için gelen bilgiyle, veri bölündükten sonra doğru sınıflama yapmak için gelen bilgi arasındaki fark kazanım olarak adlandırılır. Başlangıçtaki entropi ile her bir alt bölümün entropilerinin ağırlıklı toplamalarının farkı alınır. Farkı büyük olan alt bölüme doğru dallanma yapılır. Bilgi kazanımı Denklem 4.19'daki gibi hesaplanır [106].

$$Kazanım(D; S) = H(D) - \sum_{i=1}^n P(D_i)H(D_i) \quad (4.19)$$

4.3. Uygulama Mimarisi

Bu bölümde, önceki bölümlerde anlatılanlar ışığında toplam 3336 aminoasit uzunluğundaki yirmi iki farklı hemoglobin proteininin birincil yapısı kullanılarak ikincil yapısını tahmin edebilmek adına, sınıflandırma başarısını arttırmak için KSA ve iyileştirilmiş KSA kullanılarak gerçekleştirilen uygulama adımları paylaşılmıştır.

Protein yapı bilgileri, Protein Data Bank (PDB)'ta araştırmacılara ücretsiz olarak sunulmaktadır [107]. PDB'de proteinler, rakam ve harflerden oluşan dört basamaklı kod ile temsil edilmektedir. Çalışmada kullanılan protein verisi PDB kodları ile birlikte Tablo 4.8.'de verilmiştir.

Örneğin 1A4F PDB kodu ile sorgulama yapıldığında hemoglobin proteininin Şekil 4.7.'de verilen yapısına ulaşılabilir.

Protein dizisinde üstteki satırlar proteinin birincil yapısını, alttaki satırlar ise ikincil yapısını göstermektedir.

Tablo 4.8. Çalışmada kullanılan proteinler

PDB kodu	Kullanılan zincir
1BUV	M ve T zincirleri
1HCO	A ve B zincirleri
1C7D	A ve B zincirleri
2HHB	A ve D zincirleri
1HHO	A ve B zincirleri
2HBS	D ve E zincirleri
1HV4	D ve E zincirleri
4HHB	A ve D zincirleri
1A4F	A ve B zincirleri
1DLW	A zinciri
3GOU	A zinciri
3HRW	A zinciri
1NQP	A zinciri

1 VLSAADKTNV KGVFSKISGH AEEYGAETLE RMFTAYPQTK TYFPHFDLQH
 HHHHHHH HHHHHHHTT HHHHHHHHHH HHHHH GGGG GG TTS ST

51 GSAQIKAHGK KVVAAALVEAV NHIDDIAGAL SKLSDLHAQK LRVDPVNFKF
 T HHHHHHH HHHHHHHHH TTTT HHHHT HHHHHHHTT S THHHHH

101 LGHCFLVVVA IHHPALTAE VHASLDKFLC AVGTVLTAKY R
 HHHHHHHHH HH TTT HH HHHHHHHHH HHHHHHTT TT

Şekil 4.7. 1A4F kodlu Hemoglobin proteinin birincil ve ikincil yapısı

Veri temininden sonra, KSA ile eğitim ve sınıflandırma aşamasına geçilir. Orijinal KSA ve iyileştirilmiş KSA ile eğitim aşaması için, .NET çatısı altında C# programlama dili ile yazılım yapılmıştır. Database engine olarak MS-SQL Server kullanılmıştır.

Şekil 4.8.'de verilen form ekranında *birincil yapı* ve *ikincil yapı* alanlarına, ikincil yapısı tahmin edilmek istenen protein ya da veri setinin birincil ve ikincil yapıları girilmelidir. Birincil yapı ve ikincil yapı aynı uzunlukta olmalıdır. Birincil yapıda yirmi farklı harf, ikincil yapıda sekiz farklı harf, boşluk ya da - olmalıdır. Bu kurallara uymayan veri girişi olduğunda program uyarı vermektedir.

Form1

Birincil Yapı

KVNVDGAEALRLIVYPWTQRFFSSFGNLSPTAILGNPMVRAHGKKVLTSGDAVKNLNDNIKNTFAQLSELHCDKLHVDPENFRLLGDILIVLAHFAKEF
 TPDCQAAWQKLVVRVAHALARKYHSLFEQLGGQAAVQAVTAQFYANIQADATVATFFNGIDMPNQNTKTAFLCAALGGPNAWTGRNLKEVHANMGVNSA
 QFTTVIGHLRSLTGGVAAALVEQTVAVAETVRGDVVTVLSPADKTNIKSTWDKIGGHAGDYGGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHGKVA
 DALTTAVAHLDLPGALSALSHAYKLRVDPVNFKLLSHCLLVTLACHHPTTEFTPAVHASLDKFFAAVSTVLTSKYRVLSGEDKSNIAAWGKIGGHGAEYGA
 EALERMFASFPTTKTYFPHFDVSHGSAQVKGHGKVDALASAAGHLDLPGALSALSHAYKLRVDPVNFKLLSHCLLVTLASHHPADFTPAVHASLDKFL
 ASVSTVLTSKYRVLSPADKTNVKAAWGKVGAGHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKVDALTNVAHVDDMPNALSALSHAH
 KLRVDPVNFKLLSHCLLVTLAAHLP AEFTP AVHASLDKFLASVSTVLT SKYR

İkincil Yapı

SSHHHHHT HHHHHHHHHHHHHHHHHHTTT HHHHHHHHHHHHHHSTS THHHHHHHHHHHHHHHHGGGS HHHHHHHHHHHHHHHHHSTT
 -HHHHTTSHHHHHHHHHHHHHHHHTTT TTTGGGGTT HHHHHHHHHHHHHHHHTT SS S HHHHHTTS HHHHHHHHHHHHHHHHTT
 HHHHHHHHHHHHHHHHTTTHHHH - HHHHHHHHHHHHHHGGGHHHHHHHHHHHHHHHHH GGGGGG SSS STT HHHHHHHHHHHHHHHHTT TT
 HHHHTHHHHHHHHHTTT THHHHHHHHHHHHHHHH TTT HHHHHHHHHHHHHHHHHHTTTT-
 THHHHHHHHHHHHTGGGHHHHHHHHHHHHHHH SGGGGG SSS STT HHHHHHHHHHHHHHHHTT S TTTSTTHHHHHHTTS
 THHHHHHHHHHHHHHHHHH TTT HHHHHHHHHHHHHHHHHHGGGG- HHHHHHHHHHHHTTTGGGHHHHHHHHHHHHHHHHH GGGGGG TTS STT
 HHHHHHHHHHHHHHHHHHTTSHHHHTHHHHHHHTTS THHHHHHHHHHHHHHHH TTT HHHHHHHHHHHHHHHHHH -

Pencere Boyutu: 15 Eğitim Seti Oranı (%): 80 En Yakın Komşu Sayısı: 3
 Klon Sayısı: 1 Mutasyon Oranı: 5 Mutasyon Kolonları:
 İterasyon Sayısı: 4

İşlemlere Başla

Şekil 4.8. C# form ekranı

Verinin ne kadarının eğitim için, ne kadarının test için kullanılacağı *eğitim seti oranı* alanı ile belirlenir.

Antijenleri en iyi tanıyan antikorların kaç kez çoğaltılacağı *klon sayısı* alanı ile belirlenir.

Çoğaltma işlemi uygulanan aday antikorlar kümesinde kaç sütunda farklılaşma olacağı *mutasyon oranı* ile belirlenir.

Eğer sınıflandırma için kNN algoritması kullanılacaksa *en yakın komşu sayısı* alanına, k ile ifade edilen komşu sayısı girilmelidir. Farklı bir sınıflandırıcı kullanılacaksa KSA ile eğitim işlemi tamamlandıktan sonra program sonlandırılır.

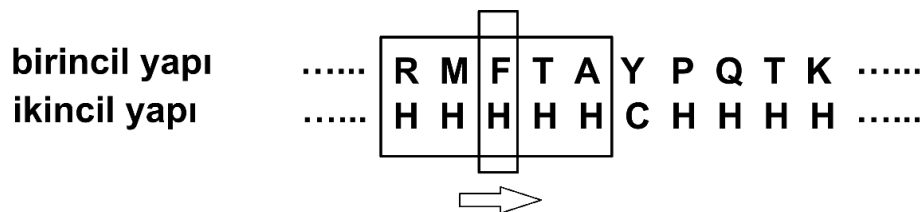
Proteinlerin ikincil yapılarının belirlenmesinde bir standart sağlamak için DSSP (Dictionary of Protein Secondary Structure) adı verilen program, Kabsch ve Sander tarafından tasarlanmıştır [108]. DSSP'ye göre proteinlerin ikincil yapıları G,H,I,T,E,B,S ve C (ya da boşluk) ile temsil edilen sekiz farklı motiften meydana gelmektedir.

- H = α -helix
- B = residue in isolated β -bridge
- E = extended strand, participates in β ladder
- G = 3-helix (3_{10} helix)
- I = 5 helix (π -helix)
- T = hydrogen bonded turn
- S = bend
- C = coil

DSSP kuralları ile bu sekiz yapı, üç temel ikincil yapıya indirgenebilmektedir. Bu indirgeme işlemi sayesinde üç sınıflı veri ile sınıflandırma çalışmaları yapmak daha kolay hale gelmektedir. Bu çalışmada kullanılan DSSP kuralına göre ikincil yapıdaki H ve G'lerin tamamı H, E ve B'lerin tamamı E ve geri kalan tüm yapılar C olarak kabul edilmiştir.

Giriş verisi uygunluğu ve DSSP indirgeme işlemleri C# platformunda yapılmıştır. Tek satır halinde olan protein verisi ile işlem yapılamayacağı için, kayan pencere yöntemi ile veri satır-sütun haline getirilmelidir. Bu işlem için tek sayı uzunluklu bir pencere boyutu seçilir; pencere, proteinin sol başından başlayarak sağa doğru birer aminoasit kayarak tarama yapar. Pencerenin taradığı her alan bir satırı oluşturur. Merkez kalıntıya karşılık gelen ikincil yapı o satırın ait olduğu sınıf olarak belirlenir. Proteinin tamamı kayan pencere ile tarandıktan sonra veriler algoritmanın uygulanması için hazır hale gelmiştir.

Şekil 4.9.'da beş pencere boyutlu kayan pencere ile protein verisinin taranma aşaması verilmiştir. Şekil 4.10.'da ise kayan pencere yönteminin uygulanması sonucu elde edilen satır-sütun formatındaki yapı verilmiştir.



Şekil 4.9. Kayan pencere yöntemi

1	2	3	4	5	Sınıf
R	M	F	T	A	H
M	F	T	A	Y	H
F	T	A	Y	P	H
T	A	Y	P	Q	C

Şekil 4.10. Kayan pencere yöntemi uygulandığında elde edilen veri formatı

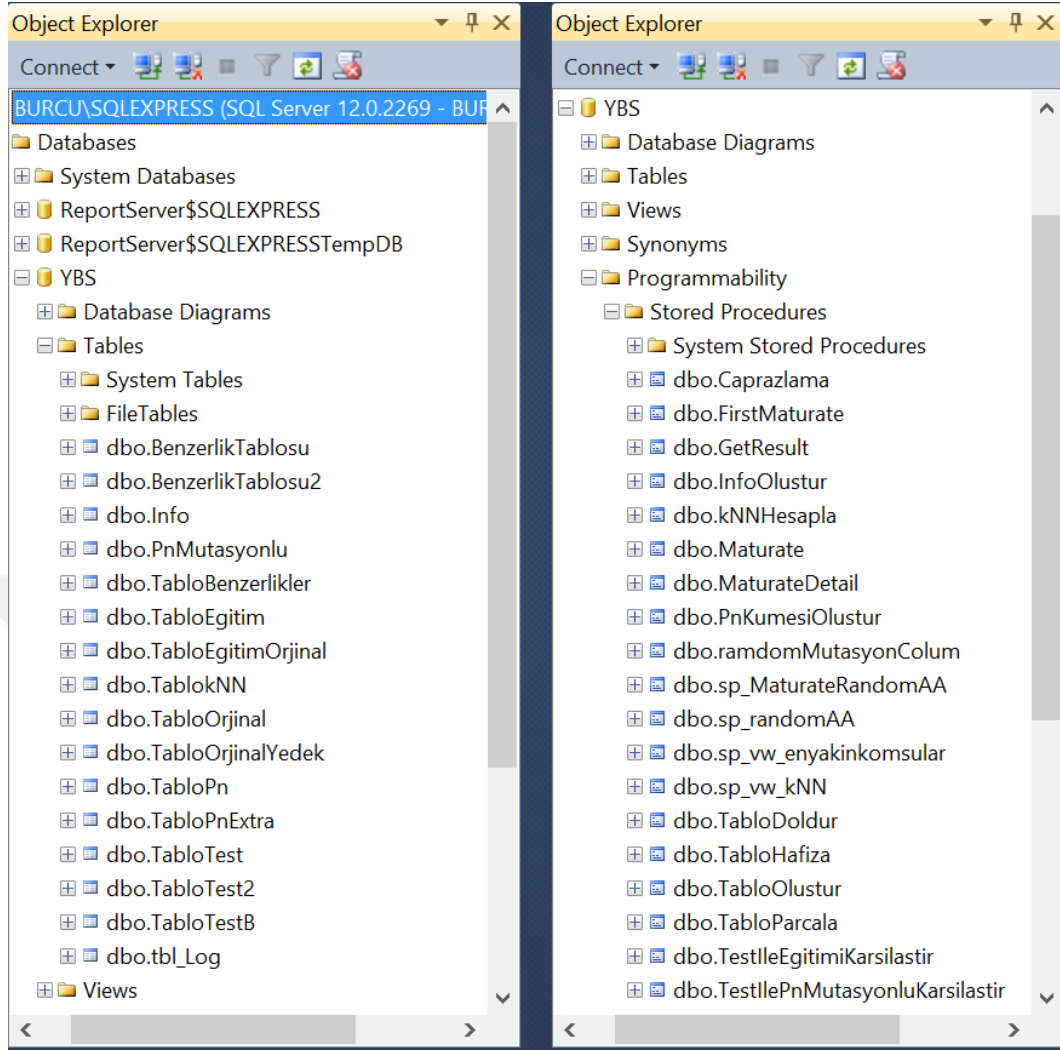
Kayan pencere ile veriyi satır-sütun formatına dönüştürme ve KSA ile ilgili işlemler saklı yordamlarla veri tabanı tarafında yapılmaktadır. Her adımda elde edilen veri ayrı tablolarda tutulmaktadır. Şekil 4.11.'de veri tabanında kullanılan tablo ve saklı yordam isimleri verilmiştir.

InfoOlustur adlı saklı yordam ile C# ortamında hazırlanan ekrandan girilen değerler programın çalışma aşamalarında kullanılmak üzere *Info* adlı tabloda tutulmaktadır.

Tek bir satır halinde bulunan protein birincil ve ikincil yapıları, *TabloOlustur* yordamı ile kayan pencere yöntemi kullanılarak satır-sütun formatına getirilip *TabloOrjinal* adlı tabloya kaydedilmektedir.

Kullanıcı tarafından girilen yüzde değerine göre *TabloOrjinal* deki veriler eğitim ve test olarak *TabloEgitim* ve *TabloTest* tablolarına kaydedilir.

Ön işlemlerden sonra protein verileri klonal seçim algoritması ile eğitmeye hazır hale gelmektedir. Algoritmanın her aşaması için ayrı saklı yordamlar hazırlanmış olup, her biri *InfoOlustur* yordamı içerisinde tetiklenmektedir.



Şekil 4.11. Veri tabanında kullanılan tablolar ve saklı yordamlar

TestIleEgitimiKarsilastir yordamı ile her bir test satırı için eğitim tablosundaki tüm satırlar arasındaki mesafe Hamming uzaklık ölçütüne göre hesaplanır ve *BenzerlikTablosuna* kaydedilir.

BenzerlikTablosundan her bir test verisi (antijen) için en benzer eğitim satırı (antikor) belirlenir ve *TabloPn*'e kaydedilir.

TabloPn klonlanır, rasgele seçilen sütunlardaki veriler kendi aralarında rasgele yer değiştirir (mutasyon).

Mutasyon işlemi sonrası elde edilmiş olan *PnMutasyonlu* tablosunun satırları ile *TabloTest* satırları Hamming uzaklık ölçütüne göre karşılaştırılarak benzerlik hesaplanır. Bu işlemler *TestilePnMutasyonluKarsilastir* yordamı içerisinde yapılır.

Bu ikinci seçme işleminde eşik değerini geçen benzerliğe sahip *PnMutasyonlu* tablosundaki satırlar hafıza hücreleri olarak belirlenmiş olur ve başlangıçtaki eğitim kümesine dahil edilir.

Bu adımlar, kullanıcı tarafından belirlenen iterasyon sayısı kadar tekrarlanır. KSA sonrasında hem veri sayısında artış olur hem de antijenleri daha iyi tanıyan antikorlara sahip olunur.

Çalışmanın sınıflandırma aşaması, ekran görüntüleri Şekil 4.12. ve Şekil 4.13.'te verilmiş olan RapidMiner Studio Educational 8.1 platformunda yapılmıştır. Rapidminer makine öğrenmesi ve veri madenciliği problemleri için geliştirilmiş bir yazılımdır. Akademisyenler ve öğrenciler ticari olmayan çalışmaları için Rapidminer'a başvurarak ücretsiz kullanım talep edebilirler.

Sınıflandırma başarısının geçerliliği on kat çapraz doğrulama yöntemi ile sınanmıştır. Veri on alt kümeye ayrılmıştır, her seferinde bir alt küme test, dokuz alt küme eğitim olacak şekilde on kez sınıflandırma yapılmıştır. Tahmin başarısı bu on sınıflandırmanın ortalama başarısı olarak kabul edilmiştir.

ÇKA ile sınıflandırma Rapidminer programının “derin öğrenme” başlığı altında yapılmaktadır. Rapidminer, açık kaynak kodlu H2O alt yapısını kullanarak derin öğrenme algoritmalarını yürütür.

Sonuçlar Q_3 başarı kriterine göre değerlendirilmiştir; doğru tahmin edilen toplam sarmal (q_H), tabaka (q_E) ve iplik (q_C) sayısının, toplam ikincil yapı sayısına oranı olarak ifade edilen eşitlik Denklem 4.20'de verilmiştir.

$$Q_3 = \frac{q_H + q_E + q_C}{n} * 100 \quad (4.20)$$

//Local Repository/deeplearning - RapidMiner Studio Educational 8.1.003 @ DESKTOP
 File Edit Process View Connections Cloud Settings Extensions Help

Views: Design Results Auto Model

Find data operators etc. All Studio Search

Operators

Search for Operators

- Data Access (47)
- Blending (77)
- Cleansing (26)
- Modeling (130)
 - Predictive (61)
 - Lazy (2)
 - Bayesian (2)
 - Trees (9)
 - Rules (5)
 - Neural Nets (4)
 - Deep Learning
 - Neural Net
 - AutoMLP
 - Perceptron
 - Functions (8)
 - Logistic Regression (3)
 - Support Vector Machines (7)
 - Discriminant Analysis (3)

[Get more operators from the Marketplace](#)

Repository

+ Add Data

- Samples
- DB
- Local Repository (sau)
- Cloud Repository (disconnected)

ALT + click to disconnect

Process

Process

100%

Read Excel

Cross Validation

The process diagram shows a flow from an input to a "Read Excel" operator, which connects to a "Cross Validation" operator. The "Cross Validation" operator has multiple output ports labeled "exa", "mod", "tes", "per", and "res". The "tes" and "per" outputs are connected to a join operator (represented by a circle with an 'X'). The "res" output is connected to a final output port.

Parameters

Cross Validation

- ☐ split on batch attribute
- ☐ leave one out
- number of folds: 10
- sampling type: automatic
- ☐ use local random seed
- ☒ enable parallel execution

[Hide advanced parameters](#)

Help

Cross Validation
Concurrency

Tags: [Cross-Validations](#), [Cross-validations](#), [Folds](#), [K-Folds](#), [K-folds](#), [Validations](#), [Estimations](#), [Evaluations](#), [Performances](#), [Splitting](#), [X-Validation](#), [X-Prediction](#), [Validation](#)

Synopsis

This Operator performs a cross validation to estimate the statistical performance of a learning model.

[Jump to Tutorial Process](#)

Description

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

☒ Activate Wisdom of Crowds

Şekil 4.12. Rapidminer ekran 1

Local Repository/deeplearning - RapidMiner Studio Educational 8.1.003 @ DESKTOP

File Edit Process View Connections Cloud Settings Extensions Help

Views: Design Results Auto Model

Find data, operators... All Studio Search

Operators

Search for Operators

- Data Access (47)
- Blending (77)
- Cleansing (26)
- Modeling (130)
 - Predictive (61)
 - Lazy (2)
 - Bayesian (2)
 - Trees (9)
 - Rules (5)
 - Neural Nets (4)
 - Deep Learning
 - Neural Net
 - AutoMLP
 - Perceptron
 - Functions (8)
 - Logistic Regression (3)
 - Support Vector Machines (7)
 - Discriminant Analysis (3)

[Get more operators from the Marketplace](#)

Repository

Add Data

- Samples
- DB
- Local Repository (sau)
- Cloud Repository (disconnected)

Process

Process Cross Validation

100%

Training

Testing

Deep Learning

Apply Model

Performance

Parameters

Deep Learning

activation: Rectifier

hidden layer sizes: Edit Enumeration (2)...

☐ reproducible (uses 1 thread)

epochs: 10.0

☐ compute variable importances

train samples per iteration: -2

☒ adaptive rate

[Hide advanced parameters](#)

[Change compatibility \(7.6.001\)](#)

Help

Deep Learning

H2O

Tags: Supervised, Classification, Regression, Model, Neural, Network, Artificial, Backpropagation, Perceptron, H2O, Neural Nets

Synopsis

Executes Deep Learning algorithm using H2O 3.8.2.6.

[Jump to Tutorial Process](#)

Description

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

☒ Activate Wisdom of Crowds

Şekil 4.13. Rapidminer ekran 2

BÖLÜM 5. UYGULAMA ÇIKTILARI

Hemoglobin proteininin ikincil yapısını tahmin etmek için KSA ile birlikte farklı sınıflandırıcılar kullanılmıştır. KSA, sınıflandırma başarısını arttırmak için önerilmiştir.

Çalışmanın ilk adımı uygun pencere büyüklüğünün belirlenmesidir. En uygun pencere büyüklüğünü belirlemek için yapılan önceki çalışmalarda, 7-17 arasındaki tek sayılardan oluşan pencere boyutları ve iterasyon sayısına bağlı sınıflandırma başarısı 1A4F Hemoglobin proteini için incelenmiş ve sonuçlar Tablo 5.1.'de verilmiştir [109].

Tablo 5.1. 1A4F proteininin farklı pencere boyutları ve iterasyonlar için sınıflandırma başarı değişimi

Pencere Boyutu	3 iterasyon	4 iterasyon	5 iterasyon	6 iterasyon	7 iterasyon
7	94	96	95	98	98
9	90	94	95	97	98
11	92	94	96	97	97
13	90	95	97	97	97
15	90	92	95	97	98
17	88	94	95	97	97

Tez çalışmasında kullanılan hemoglobin veri seti için uygun pencere boyutu belirleme çalışmasından elde edilen sonuçlar Tablo 5.2.'de verilmiştir [110]. Pencere boyutu olarak 11-21 arasındaki değerler ile deneme yapılmış ve 15 pencere boyutu tercih edildiğinde sınıflandırma başarısının daha yüksek olduğu görülmüştür. Sonuçları Tablo 5.1.'de verilmiş olan çalışma tek bir protein verisi için yapıldığından pencere boyutları daha küçük seçilmiştir. Tablo 5.2.'de sonuçları verilmiş olan çalışmada büyük veri kullanıldığından pencere boyutları daha büyük seçilmiştir.

Elde edilen sonuçlar ve literatürdeki tercih edilme sıklığı göz önüne alındığında, Hemoglobin proteini için 15 pencere boyutunun tercih edilmesi uygun görülmüştür.

Kayan pencere yöntemi ile tek satır halindeki protein verisinin tamamı soldan sağa doğru taranarak satır-sütun formatına dönüştürülmüştür. Tanımlayıcı on beş sütun ve bir hedef sütun ile sınıflandırma çalışmaları yapılmıştır.

Tablo 5.2. Pencere boyutlarına göre başarı değişimi

Pencere Boyutu	Başarı(Q ₃)
11	75.18
13	83.49
15	88.38
17	77.13
19	75.15
21	75.00

KSA'nın orijinal sınıflandırıcısı kNN olduğu için yapılan ilk çalışmalarda KSA ile birlikte kNN sınıflandırıcısı kullanılmıştır.

Farklı araştırmacılar tarafından Hemoglobin veri seti kullanarak farklı yöntemlerle yapılmış olan tahmin çalışmaları, orijinal KSA ve kNN ile yapılan ikincil yapı tahmin çalışması ile mukayese edilmiştir [110]. Elde edilen sonuçlar Tablo 5.3.'te verilmiştir.

KSA'nın orijinal versiyonunun uygulanmasıyla elde edilen olumlu sonuçlardan sonra başarıyı daha da arttırmak amacı ile algoritmada bazı değişiklikler de önerilmiştir.

Tablo 5.3. KSA ve kNN ile yapılan hemoglobin ikincil yapı tahmininin farklı yöntemler ile mukayesesi

Yöntem	Sınıflandırma Başarısı
Kendi Kendine Organize olan Global Sıralamalı Sınıflandırma Algoritması [69]	87.75
Kendi Kendine Organize olan Haritalar [69]	86.99
Genelleştirilmiş Regresyon Yapay Sinir Ağları [69]	85.93
Çok Katmanlı Algılayıcı [69]	80.94
Olasılıklı Yapay Sinir Ağları [69]	86.08
Sigmoid Kernel [111]	70.05
Polynomial Kernel [111]	73.25
Gaussian Radial Basis Kernel [111]	77.64
C-F (The Chou ve Fasman) [112]	78
GOR (The Garnier, Osguthorpe, Robson) [112]	62
PHD (The EMBL Profile Neural Network) [112]	86
PSA (The Protein Sequence Analysis) [112]	60
KSA	88.38

Önerilen iyileştirilmiş KSA yaklaşımında öncelikle verilerin eğitim ve test olarak ayrılması safhasında iki farklı yöntem denenmiştir. Elde edilen başarı yüzdeleri Tablo 5.4. ve Tablo 5.5.'te verilmiştir. Birinci yöntem verilerin başlangıçta rasgele eğitim ve test olarak ayrılması ve aynı test ile tüm KSA iterasyonlarının tamamlanmasıdır. İkincisi ise her iterasyonda farklı eğitim ve test kümeleri ile verinin eğitilmesi ve sınıflandırma için başlangıçta ayrılan bağımsız bir test kümesinin kullanılmasıdır.

Pn kümesi oluşturma aşamasında orijinal KSA'da her test satırına en yakın eğitim satırı seçilmektedir. Şekil 5.1.'de görüldüğü gibi bir test satırına en benzer olan bazı eğitim satırları diğer test satırları için bir önem teşkil etmemektedir. Toplam benzerliği çok yüksek olmasına rağmen herhangi bir test satırına en benzer olmağı için Pn kümesine seçilemeyen eğitim kümesi elemanları mevcuttur. Bu durumlar göz önünde bulundurularak, sadece tek bir test satırına benzer olduğu için değil, tüm test satırları için yüksek benzerliğe sahip olan eğitim satırlarının Pn kümesine seçiminin başarıya etkisi olup olmadığı incelenmiştir. Tüm test satırlarına olan benzerlikleri toplamı en fazla olan eğitim kümesi verileri PnExtra kümesi elemanları olarak seçilmiş ve sonuçlar değerlendirilmiştir.

s1	s2	s3	s4	s5	class	ToplamBenzerlik
T	K	T	Y	F	H	6
S	G	H	A	E	C	4
A	E	T	L	E	H	4
V	F	S	K	I	H	4
A	E	E	Y	G	H	4
A	A	D	K	T	H	4
V	K	G	V	F	H	4
H	F	D	L	Q	C	3
D	K	T	N	V	H	3
T	Y	F	P	H	C	3
G	H	A	E	E	H	3
G	A	E	T	L	H	3
D	L	Q	H	G	C	3
Y	G	A	E	T	H	3
A	Y	P	Q	T	H	3
S	A	A	D	K	H	3
L	S	A	A	D	H	3

Şekil 5.1. Her antijen için en benzer antikor ve toplam benzerlikler

KSA'da yapılan mutasyon işlemlerinin genetik algoritmalara benzerliği düşünülmüş ve mutasyon öncesi çaprazlama yapılırsa sonucu nasıl etkileyeceği incelenmiştir.

Hemoglobin A ve Hemoglobin B proteinleri için on beş pencere boyutu, 5 mutasyon oranı ile 3 iterasyon eğitim yapılmış ve eğitilmiş veri en yakın komşu 5 alınarak kNN algoritması ile sınıflandırılmıştır. Yapılan her değişiklik için bulunan başarı oranları Tablo 5.4. ve Tablo 5.5.'te verilmiştir.

Tablo 5.4. Tek test ile eğitim yapıldığında başarı

	Hemoglobin A	Hemoglobin B	Hemoglobin A+B
<i>Orijinal algoritma</i>	0,64	0,74	0,78
<i>Çaprazlama</i>	0,65	0,78	0,77
<i>PnExtra</i>	0,69	0,81	0,72
<i>Çaprazlama ve PnExtra</i>	0,69	0,72	0,72

Tablo 5.5. Her iterasyonda değişen testler ile eğitim yapıldığında başarı

	Hemoglobin A	Hemoglobin B	Hemoglobin A+B
<i>Orijinal algoritma</i>	0,63	0,8	0,72
<i>Çaprazlama</i>	0,6	0,75	0,73
<i>PnExtra</i>	0,64	0,82	0,76
<i>Çaprazlama ve PnExtra</i>	0,67	0,77	0,75

Hemoglobin A proteininde aynı test kümesi ile iterasyonlar yapıldığında, çaprazlama ve PnExtra yöntemleri sınıflandırma başarısını arttırmıştır. Her iterasyonda farklı test ile eğitim yapıldığında ve bağımsız test ile sınıflandırıldığında çaprazlama başarıyı düşürmüştür fakat PnExtra başarıyı arttırmıştır.

Hemoglobin B proteininde aynı test kümesi ile iterasyonlar yapıldığında, çaprazlama ve PnExtra yöntemleri ayrı ayrı uygulandığında sınıflandırma başarısı artmış birlikte uygulandığında azalmıştır. Her iterasyonda farklı test ile eğitim yapıldığında ve bağımsız test ile sınıflandırıldığında, Hemoglobin A da olduğu gibi çaprazlama başarıyı düşürmüştür fakat PnExtra başarıyı arttırmıştır.

Pn kümesinin oluşturulmasında PnExtra yöntemi uygulandığında sınıflandırma başarısında artış olduğu için daha sonraki çalışmalarda bu yöntem benimsenmiştir.

Bu tez çalışmasından türetilmiş kongre bildirisi ve bu bildirinin genişletilmiş indeksli yayınında KSA ile birlikte sınıflandırıcı olarak ÇKA kullanılmıştır. Elde edilen ayrıntılı sonuçlar aşağıdaki tablolarda paylaşılmıştır [113, 114].

ÇKA ile sınıflandırma aşamasında, her biri 50 nörondan oluşan gizli katmanların sayısı 1-6 arasında değiştirilerek sınıflandırma başarı değişimi incelenmiştir. Epoch sayısı 10-200 arasında değiştirilmiş, 10, 50, 80 ve 200 epoch için elde edilen sonuçlar Tablo 5.6., Tablo 5.7., Tablo 5.8. ve Tablo 5.9.'da verilmiştir.

KSA ile 5 iterasyon iyileştirme yapılmış veri ile sınıflandırma yapıldığında elde edilen sonuçlar Tablo 5.6.'da verilmiştir. 10 epoch için sınıflandırma başarısı %89,41 ile %89,96 arasında, 50 epoch için %91,74 ile %92,18 arasında, 80 epoch için %91,85 ile %92,53 arasında ve 200 epoch için %91,94 ile %92,56 arasında olduğu gözlemlenmiştir.

Tablo 5.6. KSA ile 5 iterasyon yapıldıktan sonra ÇKA ile sınıflandırma başarısı

Epoch sayısı	1 Gizli katman	2 Gizli katman	3 Gizli katman	4 Gizli katman	5 Gizli katman	6 Gizli katman
10	89,96	89,6	89,82	89,85	89,71	89,41
50	91,74	92,18	91,88	91,83	92,02	92,05
80	91,85	91,94	92,53	92,53	92,05	92,01
200	92,12	92,56	92,50	92,18	92,24	91,94

KSA ile 6 iterasyon iyileştirme yapılmış veri ile sınıflandırma yapıldığında elde edilen sonuçlar Tablo 5.7.'de verilmiştir. 10 epoch için sınıflandırma başarısı %89,92 ile %90,32 arasında, 50 epoch için %91,73 ile %92,30 arasında, 80 epoch için %91,79 ile %92,47 arasında ve 200 epoch için %91,31 ile %92,42 arasında olduğu gözlemlenmiştir.

Tablo 5.7. KSA ile 6 iterasyon yapıldıktan sonra ÇKA ile sınıflandırma başarısı

Epoch sayısı	1 Gizli katman	2 Gizli katman	3 Gizli katman	4 Gizli katman	5 Gizli katman	6 Gizli katman
10	90,16	89,92	90,08	90,32	90,18	90,26
50	91,89	91,73	92,30	91,89	92,17	92,12
80	91,79	92,47	92,47	91,97	92,28	92,25
200	92,25	92,38	92,21	92,42	92,09	91,31

KSA ile 7 iterasyon iyileştirme yapılmış veri ile sınıflandırma yapıldığında elde edilen sonuçlar Tablo 5.8.'de verilmiştir. 10 epoch için sınıflandırma başarısı %92,93 ile %93,23 arasında, 50 epoch için %94,37 ile %94,85 arasında, 80 epoch için %94,71 ile %94,96 arasında ve 200 epoch için %94,92 ile %95,06 arasında olduğu gözlemlenmiştir.

Tablo 5.8. KSA ile 7 iterasyon yapıldıktan sonra ÇKA ile sınıflandırma başarısı

Epoch sayısı	1 Gizli katman	2 Gizli katman	3 Gizli katman	4 Gizli katman	5 Gizli katman	6 Gizli katman
10	93,23	93,03	93,17	92,97	92,93	93,08
50	94,45	94,76	94,85	94,69	94,69	94,37
80	94,71	94,74	94,85	94,83	94,96	94,73
200	94,92	95,06	95,06	94,99	95,02	94,94

KSA ile 8 iterasyon iyileştirme yapılmış veri ile sınıflandırma yapıldığında elde edilen sonuçlar Tablo 5.9.'da verilmiştir. 10 epoch için sınıflandırma başarısı %94,38 ile %94,89 arasında, 50 epoch için %95,96 ile %96,36 arasında, 80 epoch için %96,15 ile %96,47 arasında ve 200 epoch için %96,36 ile %96,61 arasında olduğu gözlemlenmiştir.

Tablo 5.9. KSA ile 8 iterasyon yapıldıktan sonra ÇKA ile sınıflandırma başarısı

Epoch sayısı	1 Gizli katman	2 Gizli katman	3 Gizli katman	4 Gizli katman	5 Gizli katman	6 Gizli katman
10	94,81	94,89	94,84	94,76	94,47	94,38
50	96,02	96,21	96,24	96,36	95,96	96,02
80	96,29	96,47	96,44	96,36	96,30	96,15
200	96,36	96,61	96,46	96,49	96,47	96,45

Nöron sayısı değişiminin sınıflandırmaya etkisini incelemek için 3 gizli katmanlı ÇKA için Rectifier aktivasyon fonksiyonu ile 50, 100 ve 150 nöron kullanarak yapılan sınıflandırma sonuçları Tablo 5.10.'da verilmiştir.

Tablo 5.10. Nöron sayısı değişimine bağlı sınıflandırma başarısı

Epoch sayısı	50 nöron	100 nöron	150 nöron
10	94,84	95,23	95,38
50	96,24	96,55	96,64
80	96,44	96,75	96,73
200	96,46	96,90	96,81

Nöron sayısındaki artış işlem süresini de arttırmaktadır. Nöron sayısının 100 seçilmesi hem sınıflandırma başarısı hem de işlem süresi açısından uygundur.

Aktivasyon fonksiyonu seçiminin sınıflandırma sonucuna etkisini incelemek için her biri 100 nöron içeren 3 gizli katmanlı ÇKA yapısı rectifier, tanh ve maxout için uygulanmıştır. Elde edilen sonuçlar Tablo 5.11.'de verilmiştir.

Tablo 5.11. Aktivasyon fonksiyonu deęişimine baęlı sınıflandırma başarısı

Epoch sayısı	rectifier	tanh	maxout
80	96,75	94,84	96,54

Aktivasyon fonksiyonu olarak rectifier seçildiğinde sınıflandırma başarısı dięer aktivasyon fonksiyonlarına göre daha yüksek olmaktadır.

İyileştirilmiş KSA'nın sınıflandırma başarısına katkısını daha ayrıntılı inceleyebilmek için farklı sınıflandırıcılar ile sınıflandırma yapılmış ve sonuçlar Tablo 5.12.'de verilmiştir.

Tablo 5.12. KSA uygulanmadan ve uygulandıktan sonra yapılan sınıflandırmaların başarı oranları

İyileştirilmiş KSA İterasyon sayısı	kNN	NB	ID3	ÇKA
KSA uygulanmadı	82,17	78,35	81,57	84,01
5	90,34	86,54	85,72	92,53
6	91,91	87,14	85,45	92,47
7	94,52	90,02	93,64	94,85
8	96,18	92,84	96,18	96,44

Tablo 5.12.'de görüldüğü gibi iyileştirilmiş KSA uygulanması sınıflandırma başarısına olumlu katkı sağlamaktadır. İyileştirilmiş KSA uygulanmadan yapılan sınıflandırma başarıları, iyileştirilmiş KSA uygulanmış çalışmalara göre, dört sınıflandırma algoritması için de daha düşük olmuştur. Farklı teknik altyapıları kullanan farklı sınıflandırıcılar seçilmiş olmasına rağmen, sınıflandırma öncesinde iyileştirilmiş KSA uygulanması tüm sınıflandırma algoritmaları için tahmin başarısını arttırmıştır.

BÖLÜM 6. SONUÇLAR VE ÖNERİLER

Protein katlanmalarını doğru olarak tahmin edebilmek, tıp ve sağlık sektöründe gerek çeşitli hastalık tedavilerinde gerekse yeni ilaç tasarımında her geçen gün daha önemli bir hale gelmektedir. Protein yapılarının anlaşılması ve çözülmesi laboratuvar ortamlarında uzun süreler alan, masraflı ve zorlu süreçlerdir. Bu nedenle protein yapı tahmini biyoinformatiğin günümüzde en çok ilgi gösterilen ve sonuçları merak edilen çalışma alanlarından birine dönüşmüştür.

Yukarıda anlatılanlardan hareketle bu çalışmada hemoglobin proteinin ikincil yapı tahmini çalışılmıştır. Sınıflandırma için sinir ağlarından modellenen ÇKA, mesafeye dayalı kNN, karar ağaçlarını kullanan ID3 ve istatistiksel sınıflandırıcı NB yöntemleri kullanılmıştır.

Protein ikincil yapı tahmin başarısını arttırmak için, sınıflandırma öncesinde KSA ile eğitim yapılması önerilmiştir. Geleneksel KSA uygulanması ile ikincil yapı tahmin başarısında artış sağlandığı gözlemlenmiştir. Sonrasında, sağlanan bu başarı seviyesini daha da arttırmak için geleneksel KSA'da çeşitli iyileştirmeler yapılmıştır. Algoritmaya uygulanan bu yeniliklerle birlikte iyileştirilmiş KSA'nın tahmin başarısında geleneksel KSA'ya kıyasla artış ortaya çıkmıştır.

Hemoglobin A proteini için kNN ile yapılan sınıflandırma öncesinde geleneksel KSA uygulanırsa %64 başarı, çalışmada önerilen iyileştirilmiş KSA modeli kullanılırsa %69 başarı elde edilmektedir. Hemoglobin B proteini için geleneksel KSA uygulanırsa %74 başarı, iyileştirilmiş KSA modeli uygulanırsa %81 başarı elde edilmektedir. Hemoglobin veri seti için geleneksel KSA ile %88,38 başarı elde edilmişken çalışmada önerilen iyileştirilmiş KSA kullanıldığında başarı oranının %96,44'e ulaştığı görülmüştür.

Hemoglobin veri setinin ikincil yapı tahmini için uygun pencere boyutunu saptayabilmek için 11 – 21 arasındaki pencere boyutları ile ikincil tahmin yapı çalışmaları yapılmıştır. Tez çalışmasında kullanılan Hemoglobin veri seti için en uygun pencere boyutunun 15 olduğu tespit edilmiştir.

Farklı sınıflandırıcıların protein ikincil yapı tahminindeki etkilerini görebilmek için dört farklı teknik altyapıyı kullanan dört farklı sınıflandırıcı ile çalışılmıştır. Hemoglobin veri seti iyileştirilmiş KSA ile eğitilmeden ve eğitildikten sonra söz konusu sınıflandırıcılar ile sınıflandırılmıştır.

KSA ile eğitim yapılmadan kNN ile sınıflandırma yapıldığında %82,17, iyileştirilmiş KSA ile eğitim yapıldıktan sonra sınıflandırma yapıldığında %96,18 başarı elde edilmiştir. NB ile sınıflandırmada KSA eğitimsiz veri kullanıldığında %78,35, iyileştirilmiş KSA ile eğitilen veri kullanıldığında %92,84 başarılı tahmin yapılmıştır. ID3 ile sınıflandırma aşamasında KSA eğitimsiz veri ile %81,57, iyileştirilmiş KSA eğitilmiş veri ile %96,18 tahmin başarıları elde edilmiştir. ÇKA ile sınıflandırmada iyileştirilmiş KSA ile verinin eğitilmesi tahmin başarılarını %84,01'den %96,44'e çıkartmıştır.

Elde edilen sonuçlar göstermektedir ki sınıflandırma öncesinde verilerin çalışmada önerilen iyileştirilmiş KSA yaklaşımı ile eğitilmesi, tüm sınıflandırıcılarda sınıflandırma başarısında artış sağlamıştır.

Protein ikincil yapı tahmininde kullanılması önerilen KSA yaklaşımının, sınıflandırıcıdan bağımsız olarak başarıyı arttırdığı görülmüştür. Bu çalışmada kullanılmamış olan farklı sınıflandırıcılar için de iyileştirilmiş KSA'nın sınıflandırma başarısında artış sağlaması beklenmektedir.

Protein yapı tahmininde başarıları tespit edilmiş olan iyileştirilmiş KSA yaklaşımının farklı sınıflandırma problemlerinde ve protein dışındaki verilerle de önemli tahmin yüzdelere ulaşması olasıdır.

Bu tez çalışmasında elde edilen sonuçlar ışığında herhangi bir veri kümesinin sınıflandırılmasında daha başarılı sonuçlar elde edilmesi için verinin sınıflandırmadan önce çalışmada önerilen iyileştirilmiş KSA yaklaşımı ile eğitilmesi tavsiye edilmektedir.

Bu çalışma hemoglobin veri seti kullanılarak yapılmıştır. Sonraki çalışmalarda RS126, CB513, CASP ve EVA gibi büyük veri setlerinin kullanılması önerilmektedir.

Büyük veri setleri ile çalışmak önerilen çözüm yönteminin literatürde kabul görme olasılığını arttırmaktadır. Fakat büyük veri ile çalışmak işlem karmaşıklığını da beraberinde getirmektedir. Büyük protein veri setleri ile ikincil yapı tahmini çalışmalarında derin öğrenme yöntemlerinin kullanılması bu işlem karmaşıklığının üstesinden gelinmesinde avantaj sağlayacaktır. İleriki çalışmalarda protein ikincil yapı tahmini için büyük veri setleri ile birlikte derin öğrenme yöntemlerinin kullanılması önerilmektedir.

KAYNAKLAR

- [1] Allison, L.A., Fundamental molecular biology. 2. Baskı, Wiley, 2012.
- [2] Lodish, H.F., Molecular cell biology. 7. Baskı, W.H. Freeman and Co, New York, 2013.
- [3] Reynaud, E., Protein Misfolding and Degenerative Diseases. Nat Educ 3:28, 2010.
- [4] Pauling, L., Corey, R.B., Branson, H.R., The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A 37:205–11, 1951.
- [5] Pauling, L., Corey, R.B., Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets. Proc Natl Acad Sci U S A 37:729–40, 1951.
- [6] Perutz, M.F., Rossmann, M.G., Cullis, A.F., Muirhead, H., Will, G., North, A.C.T., Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. Nature 185:416-422, 1960.
- [7] Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., Phillips, D.C., A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. Nature 181:662-666, 1958.
- [8] Rost, B., Review: Protein Secondary Structure Prediction Continues to Rise. J Struct Biol 134:204–218, 2001.
- [9] Nagano, K., Logical analysis of the mechanism of protein folding: I. Predictions of helices, loops and β -structures from primary structure. J Mol Biol 75:401–420, 1973.
- [10] Chou, P.Y., Fasman, G.D., Conformational parameters for amino acids in helical, β -sheet, and random coil regions calculated from proteins. Biochemistry 13:211–222, 1974.
- [11] Chou, P.Y., Fasman, G.D., Prediction of protein conformation. Biochemistry 13:222–245, 1974.

- [12] Lim, V.I., Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J Mol Biol* 88:857–872, 1974.
- [13] Lim, V.I., Algorithms for prediction of α -helical and β -structural regions in globular proteins. *J Mol Biol* 88:873–894, 1974.
- [14] Pirovano, W., Heringa, J., Protein Secondary Structure Prediction. İçinde: *Data Mining Techniques for the Life Sciences. Methods in molecular biology.* Humana, 327–348, 2010.
- [15] Garnier, J., Osguthorpe, D.J., Robson, B., Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* 120:97–120, 1978.
- [16] Mount, D., *Bioinformatics: Sequence & Genome Analysis.* Cold Spring Harbor Laboratory Yayınları, New York, 2001.
- [17] Zvelebil, M.J., Barton, G.J., Taylor, W.R., Sternberg, M.J.E., Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J Mol Biol* 195:957–961, 1987.
- [18] Rost, B., Sander, C., Prediction of Protein Secondary Structure at Better than 70% Accuracy. *J Mol Biol* 232:584–599, 1993.
- [19] Salamov, A.A., Solovyev, V.V., Prediction of Protein Secondary Structure by Combining Nearest-neighbor Algorithms and Multiple Sequence Alignments. *J Mol Biol* 247:11–15, 1995.
- [20] Jones, D.T., Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 292:195–202, 1999.
- [21] Cuff, J.A., Barton, G.J., Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins Struct Funct Genet* 40:502–511, 2000.
- [22] Magnan, C.N., Baldi, P., SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics* 30:2592–2597, 2014.
- [23] Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., Tramontano, A., Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins Struct Funct Bioinforma* 84:4–14, 2016.
- [24] Won, K.J., Hamelryck, T., Prügél-Bennett, A., Krogh, A., An evolutionary method for learning HMM structure: prediction of protein secondary structure. *BMC Bioinformatics* 8:357, 2007.

- [25] Ding, Y.S., Zhang, T.L., Gu, Q., Zhao, P.Y., Chou, K.C., Using Maximum Entropy Model to Predict Protein Secondary Structure with Single Sequence. *Protein Pept Lett* 16:552–560, 2009.
- [26] Liang, K., Wang, X., Protein Secondary Structure Prediction using Deterministic Sequential Sampling. *J Data Mining Genomics Proteomics* 02:1–7, 2011.
- [27] Wu, Z.Y., Han, R.P.S., SAAS: Short Amino Acid Sequence - A Promising Protein Secondary Structure Prediction Method of Single Sequence. *Int J Bioautomation* 17:65–72, 2013.
- [28] Aydin, Z., Altunbasak, Y., Borodovsky, M., Protein secondary structure prediction for a single-sequence using hidden semi-Markov models. *BMC Bioinformatics* 7:178, 2006.
- [29] Aydin, Z., Altunbasak, Y., Pakatci, I.K., Erdogan, H., Training Set Reduction Methods for Protein Secondary Structure Prediction in Single-Sequence Condition. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE*, 5025–5028, 2007.
- [30] Qian, N., Sejnowski, T.J., Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol* 202:865–84, 1988.
- [31] Holley, L.H., Karplus, M., Protein secondary structure prediction with a neural network. *Proc Natl Acad Sci U S A* 86:152–6, 1989.
- [32] Kneller, D.G., Cohen, F.E., Langridge, R., Improvements in protein secondary structure prediction by an enhanced neural network. *J Mol Biol* 214:171–182, 1990.
- [33] Rost, B., Sander, C., Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins Struct Funct Genet* 19:55–72, 1994.
- [34] Kaur, S., Pandey, B., Protein Secondary Structure Prediction Using Feed Forward Artificial Neural Network and Perceptron. *Int J Control Theory Appl* 9:974–5572, 2016.
- [35] Drozdetskiy, A., Cole, C., Procter, J., Barton, G.J., JPred4: a protein secondary structure prediction server. *Nucleic Acids Res* 43:389–394, 2015.
- [36] Asai, K., Hayamizu, S., Handa, K., Prediction of protein secondary structure by the hidden Markov model. *Comput Appl Biosci* 9:141–6, 1993.
- [37] Karplus, K., Sjölinder, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L., Sander, C., Predicting protein structure using hidden Markov models. *Proteins Struct Funct Bioinforma* 29:134–139, 1997.

- [38] Martin, J., Gibrat, J.F., Rodolphe, F., Analysis of an optimal hidden Markov model for secondary structure prediction. *BMC Struct Biol* 6:25, 2006.
- [39] Xia, X., Hidden Markov Models and Protein Secondary Structure Prediction. In: *Bioinformatics and the Cell*. Springer International Publishing, Cham, 145–172, 2018.
- [40] Ward, J.J., McGuffin, L.J., Buxton, B.F., Jones, D.T., Secondary structure prediction with support vector machines. *Bioinformatics* 19:1650–5, 2003.
- [41] Shah, H., Protein Secondary Structure Prediction Using Support Vector Machines (SVMs). *International Conference on Machine Intelligence and Research Advancement*. IEEE, 594–598, 2013.
- [42] Li, B., Li, Y., Gong, L., Protein secondary structure optimization using an improved artificial bee colony algorithm based on AB off-lattice model. *Eng Appl Artif Intell* 27:70–79, 2014.
- [43] Shmygelska, A., Hoos, H.H., An Improved Ant Colony Optimisation Algorithm for the 2D HP Protein Folding Problem. Springer, Berlin, Heidelberg, 400–417, 2003.
- [44] Ghosh, A., Parai, B., Protein secondary structure prediction using distance based classifiers. *Int J Approx Reason* 47:37–44, 2008.
- [45] Liu, T., Zheng, X., Wang, J., Prediction of protein structural class using a complexity-based distance measure. *Amino Acids* 38:721–728, 2010.
- [46] Pedersen, J.T., Moult, J., Genetic algorithms for protein structure prediction. *Curr Opin Struct Biol* 6:227–231, 1996.
- [47] Unger, R., The Genetic Algorithm Approach to Protein Structure Prediction. Springer, Berlin, Heidelberg, 153–175, 2004.
- [48] Zhang, X., Mesirov, J.P., Waltz, D.L., Hybrid system for protein secondary structure prediction. *J Mol Biol* 225:1049–63, 1992.
- [49] Wu, K.P., Lin, H.N., Chang, J.M., Sung, T.Y., Hsu, W.L., HYPROSP: a hybrid protein secondary structure prediction algorithm-a knowledge-based approach. *Nucleic Acids Res* 32:5059–5065, 2004.
- [50] Lin, H.N., Chang, J.M., Wu, K.P., Sung, T.Y., Hsu, W.L., HYPROSP II-A knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics* 21:3227–3233, 2005.
- [51] Patel, M.S., Mazumdar, H.S., Knowledge base and neural network approach for protein secondary structure prediction. *J Theor Biol* 361:182–189, 2014.

- [52] Bouziane, H., Messabih, B., Chouarfia, A., Effect of simple ensemble methods on protein secondary structure prediction. *Soft Comput* 19:1663–1678, 2015.
- [53] Hasic, H., Buza, E., Akagic, A., A hybrid method for prediction of protein secondary structure based on multiple artificial neural networks. *International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 1195–1200, 2017.
- [54] Paliwal, K., Lyons, J., Heffernan, R., A Short Review of Deep Learning Neural Networks in Protein Structure Prediction Problems. *Adv Tech Biol Med* 03:1–2, 2015.
- [55] Mamoshina, P., Vieira, A., Putin, E., Zhavoronkov, A., Applications of Deep Learning in Biomedicine. *Mol Pharm* 13:1445–1454, 2016.
- [56] Min, S., Lee, B., Yoon, S., Deep learning in bioinformatics. *Brief Bioinform* 18:851–869, 2016..
- [57] Goh, G.B., Hodas, N.O., Vishnu, A., Deep learning for computational chemistry. *J Comput Chem* 38:1291–1307, 2017.
- [58] Pollastri, G., Przybylski, D., Rost, B., Baldi, P., Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins Struct Funct Genet* 47:228–235, 2002.
- [59] Spencer, M., Eickholt, J., Cheng, J., A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction. *IEEE/ACM Trans Comput Biol Bioinforma* 12:103–12, 2015.
- [60] Li, Z., Yu, Y., Protein Secondary Structure Prediction Using Cascaded Convolutional and Recurrent Neural Networks. *International Joint Conference on Artificial Intelligence (IJCAI)*. New York City, 2560–2567, 2016.
- [61] Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y., Zhou, Y., Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep* 5:11476, 2015.
- [62] Yang, Y., Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Zhou, Y., SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural Networks. *Methods in molecular biology*, 55–63, 2017.
- [63] Zhou, J., Troyanskaya, O.G., Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. *International Conference on International Conference on Machine Learning Bildiriler Kitabı*, 32:745–753, 2014.

- [64] Wang, S., Peng, J., Ma, J., Xu, J., Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci Rep* 6:18962, 2016.
- [65] Wang, Y., Mao, H., Yi, Z., Protein secondary structure prediction by using deep learning method. *Knowledge-Based Syst* 118:115–123, 2017.
- [66] Fang, C., Shang, Y., Xu, D., MUFOLD-SS: New deep inception-inside-inception networks for protein secondary structure prediction. *Proteins Struct Funct Bioinforma* 86:592–598, 2018.
- [67] İbrikçi, T., Neural Network Models For Protein Structure Prediction. Çukurova Üniversitesi, Fen Bilimleri Enstitüsü, Elektrik ve Elektronik Mühendisliği ABD, Doktora Tezi, 2000.
- [68] Ersöz, İ., Secondary Structure Prediction of Hemoglobin by Using Combined Neural Networks. Çukurova Üniversitesi, Fen Bilimleri Enstitüsü, Elektrik ve Elektronik Mühendisliği ABD, Yüksek Lisans Tezi, 2003.
- [69] Atar, E., Yapay sinir ağları ile proteinlerin ikincil yapılarının kestirimi. Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Elektronik ve Haberleşme Mühendisliği ABD, Yüksek Lisans Tezi, 2005.
- [70] Kurt, V., Protein Structure Prediction using Decision List. Koç Üniversitesi, Fen Bilimleri ve Mühendislik Enstitüsü, Bilgisayar Bilimleri ve Mühendisliği Bölümü, Yüksek Lisans Tezi, 2005.
- [71] Demir, M., Peptit Bağ Açılarını Kullanarak Yapay Sinir Ağı Tabanlı Proteinlerin Sekonder Yapı Tahmini. Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği ABD, Yüksek Lisans Tezi, 2006.
- [72] Arslan Tuncer, S., Fidan Gelişimi Algoritması ile Protein Yapı Tahmini. Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği ABD, Yüksek Lisans Tezi, 2008.
- [73] Pakatçı, İ.K., Protein Secondary Structure Prediction With Classifier Fusion. Sabancı Üniversitesi, Mühendislik ve Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2008.
- [74] Batbat, T., Protein Yapısının Yapay Arı Kolonisi Algoritması ile Tahmini. Erciyes Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği ABD, Yüksek Lisans Tezi, 2014.
- [75] Akçeşme, F.B., Avicenna: Sequence Segments Similarity Based Protein Secondary Structure Prediction Method. Uluslararası Saraybosna Üniversitesi, Genetik ve Biyomühendislik Bölümü, Doktora Tezi, 2016.

- [76] Uzut, Ö.G., Optimizing Classifiers For Protein Secondary Structure Prediction. Abdullah Gül Üniversitesi, Fen Bilimleri Enstitüsü, Elektrik ve Bilgisayar Mühendisliği ABD, Yüksek Lisans Tezi, 2017.
- [77] Ahmad, F., Mat Isa, N.A., Hussain, Z., Osman, M.K., Intelligent Medical Disease Diagnosis Using Improved Hybrid Genetic Algorithm - Multilayer Perceptron Network. *J Med Syst* 37:9934, 2013.
- [78] Ozkan, O., Yildiz, M., Arslan, E., Yildiz, S., Bilgin, S., Akkus, S., Koyuncuoglu, H.R., Koklukaya, E., A Study on the Effects of Sympathetic Skin Response Parameters in Diagnosis of Fibromyalgia Using Artificial Neural Networks. *J Med Syst* 40:54, 2016.
- [79] Nalluri, M.R., Krithivasan, K, M. Manisha, Roy, D.S., Hybrid Disease Diagnosis Using Multiobjective Optimization with Evolutionary Parameter Optimization. *J Healthc Eng* 2017:5907264, 2017.
- [80] Moreira, M.W.L., Rodrigues, J.J.P.C., Kumar, N., Nature-Inspired Algorithm for Training Multilayer Perceptron Networks in e-health Environments for High-Risk Pregnancy Care. *J Med Syst* 42:51, 2018.
- [81] Dasgupta, D., An Overview of Artificial Immune Systems and Their Applications. *Artificial Immune Systems and Their Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, 3–21, 1993.
- [82] De Castro, L., José, F., von Zuben, A.A., *Artificial Immune Systems: Part I- Basic Theory and Applications*, FEEC/UNICAMP, Campinas, Brazilya, Teknik Rapor, 2000.
- [83] Harmer, P.K., Williams, P.D., Gunsch, G.H., Lamont, G.B., An artificial immune system architecture for computer security applications. *IEEE Trans Evol Comput* 6:252–280, 2002.
- [84] Bernardino, H.S., Barbosa, H.J.C., *Artificial Immune Systems for Optimization*. Springer, Berlin, Heidelberg, 389–411, 2009.
- [85] Timmis, J., Ismail, A.R., Bjercknes, J.D., Winfield, A.F.T., An immune-inspired swarm aggregation algorithm for self-healing swarm robotic systems. *Biosystems* 146:60–76, 2016.
- [86] Liang, C., Peng, L., An Automated Diagnosis System of Liver Disease using Artificial Immune and Genetic Algorithms. *J Med Syst* 37:9932, 2013.
- [87] Singh, R., Sengupta, R.N., Bankruptcy Prediction Using Artificial Immune Systems. Springer, Berlin, Heidelberg, 131–141, 2007.

- [88] Cutello, V., Narzisi, G., Nicosia, G., A Class of Pareto Archived Evolution Strategy Algorithms Using Immune Inspired Operators for Ab-Initio Protein Structure Prediction. Springer, Berlin, Heidelberg, 54–63, 2005.
- [89] Cutello, V., Nicosia, G., Pavone, M., Timmis, J., An Immune Algorithm for Protein Structure Prediction on Lattice Models. IEEE Trans Evol Comput 11:101–117, 2007.
- [90] Cutello, V., Morelli, G., Nicosia, G., Pavone, M., Scollo, G., On discrete models and immunological algorithms for protein structure prediction. Nat Comput 10:91–102, 2011.
- [91] Ünal, M., Hücre biyolojisi. Marmara Üniversitesi Yayınevi, 2012.
- [92] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., Hücrenin moleküler biyolojisi. TÜBA, 2008.
- [93] Nelson, D.L., Cox, M.M., Lehninger Biyokimyanın İlkeleri. Çev. Elçin, Y. M., 5. Baskı, Palme Yayınları, Ankara, 2013.
- [94] https://tr.wikipedia.org/wiki/Dosya:Main_protein_structure_levels_en.svg, Erişim Tarihi: 06.11.2018
- [95] Duygu, A., Dr. Amino Asit. İstanbul Tıp Kitabevleri, İstanbul, 2016.
- [96] <https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/orders-of-protein-structure>, Erişim Tarihi: 06.11.2018.
- [97] <https://www.scientificamerican.com/article/are-prions-behind-all-neurodegenerative-diseases1/>, Erişim Tarihi: 12.11.2018.
- [98] <http://www.focusdergisi.com.tr/saglik/00156/>, Erişim Tarihi: 12.11.2018.
- [99] <https://www.scienceinschool.org/tr/2010/issue15/prions>, Erişim Tarihi: 12.11.2018.
- [100] Aickelin, U., Dasgupta, D., Artificial Immune Systems. Search Methodologies. Springer US, Boston, MA, 375–399, 2005.
- [101] De Castro, L.N., Von Zuben, F.J., The Clonal Selection Algorithm with Engineering Applications. GECCO 2002 Bildiriler kitabı, 36-37, 2002.
- [102] Öztemel, E., Yapay sinir ağları. Papatya Yayıncılık Eğitim, İstanbul, 2016.
- [103] Rumelhart, D.E., Hinton, G.E., Williams, R.J., Learning representations by back-propagating errors. Nature 323:533–536, 1986.

- [104] Özkan, Y., Veri Madenciliği Yöntemleri. Papatya Yayıncılık Eğitim, İstanbul, 2013.
- [105] Quinlan, J.R., Induction of decision trees. Mach Learn 1:81–106, 1986.
- [106] Silahtaroglu, G., Veri Madenciliği: Kavram ve Algoritmaları. Papatya Yayıncılık, İstanbul, 2013.
- [107] <https://www.rcsb.org/pdb/explore/remediatedSequence.do?structureId=1A4F.>, Erişim Tarihi: 21.11.2018.
- [108] Kabsch, W., Sander, C., Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22:2577–2637, 1983.
- [109] Çarklı Yavuz, B., Yurtay, N., Prediction of Secondary Structures of Proteins Using Clonal Selection Algorithm. International Journal of Arts & Sciences, Prag, 285–290, 2015.
- [110] Çarklı Yavuz, B., Sertkaya, C., Yurtay, N., Prediction of secondary structures of hemoglobin using clonal selection algorithm. Int Work Comput Sci Eng, Çin, 1387–1391, 2017.
- [111] Ibrikci, T., Cakmak, A., Ersoz, I., Ersoy, O.K., Hemoglobin Secondary Structure Prediction with Four Kernels on Support Vector Machines. ICSC Congress on Computational Intelligence Methods and Applications, IEEE, 1–3, 2005.
- [112] Krystek, S.R., Metzler, W.J., Novotny, J., Protein Secondary Structure Prediction. İçinde: Current Protocols in Protein Science. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2.3.1-2.3.20, 1995.
- [113] Çarklı Yavuz, B., Yurtay, N., Ozkan, O., Çok Katmanlı İleri Beslemeli Yapay Sinir Ağları İle Hemoglobin Proteininin İkincil Yapısının Tahmini. 3. Uluslararası Mühendislik Mimarlık ve Tasarım Kongresi, Kocaeli, 65–66, 2018.
- [114] Çarklı Yavuz, B., Yurtay, N., Ozkan, O., Prediction of Protein Secondary Structure With Clonal Selection Algorithm and Multilayer Perceptron. IEEE Access, 6:45256–45261, 2018.

ÖZGEÇMİŞ

Burcu Çarklı Yavuz, 29.06.1983'te Sakarya'da doğdu. İlk ve orta öğrenimini Sakarya'da, lise eğitimini İstanbul'da tamamladı. 2001 yılında Küçükalyalı Kadir Has Lisesi'nden mezun oldu. Aynı yıl başladığı İstanbul Kültür Üniversitesi Matematik-Bilgisayar (Burslu) Bölümü'nü 2005 yılında bitirdi. 2010 yılında Sakarya Üniversitesi Bilgisayar ve Bilişim Mühendisliği Bölümü'nde yüksek lisans eğitimini tamamladı. 2013 yılından beri Sakarya Üniversitesi Bilişim Sistemleri Mühendisliği bölümünde Araştırma Görevlisi olarak görev yapmaktadır.