YALOVA ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

PROTEİNLERİN DÜZENSİZ BÖLGELERİNİN TAHMİNİNDE YENİ ÖZNİTELİK KODLAMA YÖNTEMLERİ

YÜKSEK LİSANS TEZİ

Sebahattin BABUR

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

HAZIRAN 2016

YALOVA ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

PROTEİNLERİN DÜZENSİZ BÖLGELERİNİN TAHMİNİNDE YENİ ÖZNİTELİK KODLAMA YÖNTEMLERİ

YÜKSEK LİSANS TEZİ

Sebahattin BABUR
(115105015)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Doç. Dr. Murat GÖK

HAZIRAN 2016

YALOVA Üniversitesi Fen Bilimleri Enstitüsü'nün 115105015 numaralı Yüksek Lisans Öğrencisi **Sebahattin BABUR**, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **" PROTEİNLERİN DÜZENSİZ BÖLGELERİNİN TAHMİNİNDE YENİ ÖZNİTELİK KODLAMA YÖNTEMLERİ "** başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmustur.

Tez Danışmanı : **Doç. Dr. Murat GÖK** .....
Yalova Üniversitesi

Jüri Üyeleri : **Doç. Dr. Murat GÖK** .....
Yalova Üniversitesi

Yrd. Doç. Dr. Osman Hilmi KOÇAL .....
Yalova Üniversitesi

Doç. Dr. Ecir Uğur KÜÇÜKSİLLE .....
Süleyman Demirel Üniversitesi

Teslim Tarihi : **6 Mayıs 2016**
Savunma Tarihi : **2 Haziran 2016**

Anneme ve Babama,

ÖNSÖZ

Yüksek Lisans öğrenim sürem boyunca bana yol gösteren danışmanım Sayın Doç. Dr. Murat GÖK' e, çalışmam sırasında bana fikir veren ve her alanda desteğini esirgemeyen değerli arkadaşım Uğur TURHAL' a ve tüm eğitim hayatım boyunca, maddi ve manevi her türlü desteği sağlayan aileme sonsuz teşekkürlerimi sunarım.

Haziran 2016 Sebahattin Babur

İÇİNDEKİLER

Sayfa

KISALTMALAR

FD : Fourier Dönüşümü
DVM : Destek Vektör Makineleri
PSM : Pozisyon Skor Matrisi
MKK : Matthew's Korelasyon Katsayısı
NMR : Nuclear Magnetic Resonance Spectroscopy
DA : Doğru Artılar
DE : Doğru Eksiler
YA : Yanlış Artılar
YE : Yanlış Eksiler
AUC : Area Under ROC Curve
ROC : Receiver Operating Characteic
k-EYK : k En Yakın Komşuluk

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

PROTEİNLERİN DÜZENSİZ BÖLGELERİNİN TAHMİNİNDE YENİ ÖZNİTELİK KODLAMA YÖNTEMLERİ

ÖZET

Canlıların temel yapıtaşlarından olan proteinler, biyokimyasal aktivitelerin hemen hemen tümünde rol alırlar. Biyokimyasal aktivitelerdeki değişimler canlı için kritik öneme sahiptir. Örneğin, protein işlev bozukluğundan kaynaklanan biyokimyasal reaksiyon değişimleri ciddi bir takım hastalıklara neden olabilmektedir. Bu nedenle proteinlerin yapısının tespit edilmesi hayati önem arz etmektedir. Proteinlerin işlevlerini anlayabilmek için ilk olarak yapılarının anlaşılması gerekmektedir. Bu yapıları tespit etmek amacıyla, Nuclear Magnetic Resonance Spectroscopy (NMR), Nuclear Overhauser Effect, X-Ray Crystallography, DNA microarray teknolojisi gibi in vitro (laboratuvar ortamı) yöntemler kullanılmaktadır. In silico (bilgisayar, hesaplamalı ortamı) yöntemlerin uygulaması in vitro yöntemlere göre maliyet / fayda ve zaman açısından daha çok tercih edilmektedir. Bu durum beraberinde aminoasitlerin birbirleri ile olan fizikokimyasal ilişkilerinin tespit edilmesi, proteinlerin yapılarının modellenmesi, metabolik yolların tahmin edilmesi gibi çalışmaları hızlandırmıştır.

Bu tez çalışmasında, proteinlerdeki düzensiz bölgelerin tahmini için proteini oluşturan amino asitlerin (kalıntı) farklı biyokimyasal ve fiziksel özellikleri kullanılarak yeni iki adet öznitelik kodlama yöntemi geliştirilmiştir. Geliştirilen birinci yöntem proteinlerin amino asitlerin fizikokimyasal özellikleri, proteinlerin pozisyon-skor matrisi (PSM) ve SPINE-X protein özellikleri temelinde geliştirilmiştir. İkinci yöntemin geliştirilmesinde ise, dalgacık teoremi kullanılmıştır. Bu öznitelik kodlama yöntemleri makine öğrenmesi algoritmaları ile Protein Structure Prediction Center tarafından yayınlanan CASP 10 veri seti üzerinde test edilmişlerdir. Elde edilen deneysel sonuçlar literatürdeki yöntemlerle kıyaslanmıştır.

NEW FEATURE CODING METHODS IN DISORDER REGION ESTIMATE OF PROTEIN

SUMMARY

The proteins which are the main constituent of bios (living creatures), take part in almost every biochemical activities. For instance, changes of biochemical reaction, derived from protein dysfunction, causes a set of serious illnesses. Therefore determining the structure of proteins is of vital importance. Firstly it must be understood the structure of proteins to understand the fuctions of them. To determine these structures, it is used such in vitro (laboratory environment) methods like Nuclear Magnetic Resonance Spectroscopy (NMR), Nuclear Overhauser Effect, X-Ray Crystallography, DNA microarray technology. Application of in silico (computer, calculating environment) methods are more favored over in vitro methods in terms of cost/benefit and time. This is accelerated the studies like determination of the physicochemical relations of amino acids, modeling of proteins's structures, prediction of pathways.

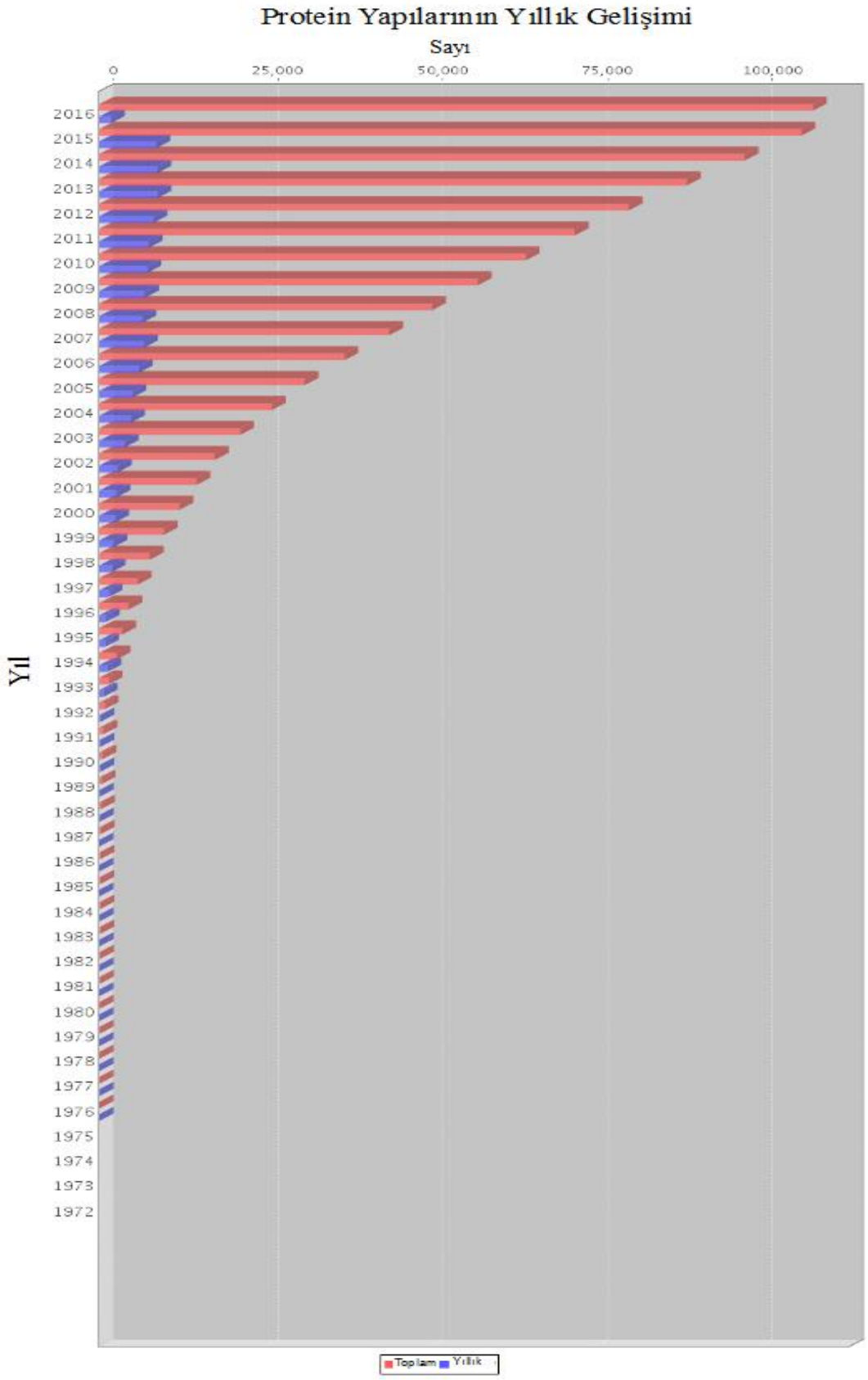
In this study it has been developed two pieces new feature codification methods by using different biochemical and physical features, of protein builder amino acids (resudiue) to predict the irregular zones of proteins. First method has been developed on the basis of physicochemical features of proteins and amino acids, position-score matris of proteins (PSM) and SPINE-X protein features. It has been used the theory of ripple in developing of second method. This feature codification methods has been tested on CASP 10 dataset which is issued by machine learning algorithms and Protein Structure Prediction Center. Acquired experimental results have been compared with the methods in the literature.

1. GİRİŞ

Proteinler yaşamın bir parçasıdır. Bu nedenle, herhangi bir canlının sahip olduğu proteinlerin çoğu enzimlerdir. Enzimler, hücre içerisindeki kimyasal reaksiyonları hızlandıran belirli proteinlerdir. Hücre içerisinde, belli bir reaksiyonun oluşabilmesi için kullanılan küçük moleküler araçlardır. Geçici olarak bileşenlere eklenirler ve bileşenlerin belli açıları yakalayabilmesini sağlarlar. Böylece istenilen reaksiyon gerçekleşebilir(Polatkan, 2007).

Proteinler, hücre içerisindeki görevlerine göre kategorilere ayrılabilirler. Birkaç kategoriye örnek vermek gerekirse; enzimler, yapısal proteinler, hormonlar, nakil proteinleri. Proteinlerin bu kadar önemli olması, proteinin yapısı hakkında daha çok bilgi edinilmesini zorunlu hale getirmiştir. Bu nedenle, proteinin üst seviye yapıları, o proteinin fonksiyonlarını, faaliyetlerini ve çevre ile olan etkileşimlerini tanımlamada önemli olmasını sağlamaktadır.

Son yıllarda, yaklaşık bir milyon protein tanımlanmıştır. Ancak, tanımlanan proteinler arasından 100.000 protein yapısı tespit edilebilmiştir. Şekil 1.1'de, protein veri bankası tarafından yıllık ve toplam tespit edilen protein yapılarının yıllara göre gelişimi gösterilmiştir(RCSB Protein Data Bank, 2016).



Şekil 1.1: PDB Veri bankasının 1972-2016 yılları arasındaki üstel büyüme.

Yüksek miktarda verilerin ortaya çıkmasıyla birlikte bu verilerin analiz edilmesi güç hale gelmiştir. Son yıllardaki istatiksel modelleme ve sınıflandırma algoritmalarındaki araştırmalar bu verilerin analizinde yeni yöntemlerin gelişmesini sağlamıştır. Bilgi teknolojisi de hızla gelişmiş ve bu büyük verileri işlemek için yöntemler geliştirilmesini sağlamıştır (Kaya, 2008).

Protein dizilerinin ve sıralı verilerin çok olması aminoasit dizilerinden protein yapısını tahminine yönelik çalışmaları motive etmiştir. Bu tahminler daha sonra düzensiz protein yapılarını anlamak için temel oluşturmuştur (Fischer, 1984).

1.1 Tezin Amacı

Bu çalışmada proteinlerin düzensiz bölgelerinin tespitine yönelik istatistiksel yöntemlerin yanında,yeni öznitelik uzayı önerilerek düzenli ve düzensiz bölgeler arasındaki en iyi sınıfsal ayrım yapılması amaçlanmıştır. Problemin çözümüne yönelik literatürde geliştirilen ve tercih edilen yöntemler kullanılarak birbirleri ile kıyaslanmıştır.

Sınıflandırıcılar için, proteinlerdeki her bir aminoasit aslında bir örnek olarak düşünüldüğünden, bunu en iyi öznitelik uzayı ile temsil etmek iyi bir sınıflandırıcı kadar önemlidir.

Önerilen yöntemin literatürde tercih edilen öznitelik uzayı oluşturma yöntemlerine alternatif olması amaçlanmaktadır. Bu çalışmanın temelini oluşturan, protein dizileri ve genom araştırmalarında son zamanlarda istatistiksel yöntemlerin yanında biyolojik verilerin de ayrık zamanlı sistemler gibi düşünülebileceğini kanıtlayan uygulamalar yapılmıştır. Bu tür ayrık zaman diferansiyel sistemlerin, biyolojik dizilerde kullanılması, genom araştırmalarında yeni bir yöntemin önünü açmıştır. Son zamanlarda, DNA araştırmalarında fourier dönüşüm temelli dalgacık yönteminin kullanılmış olması protein dizilerinde de düzensiz bölgelerin tespitinde yeni özellik vektörlerinin dalgacık yöntemiyle tespit edilebileceği fikrini ön plana getirmiştir.

Makine öğrenmesi yöntemleri ile geliştirilen araçların belirli bir başarı düzeyinden öteye geçememiş olmasından dolayı, bu çalışmada yeni öznitelik oluşturma yöntemlerinin araştırılması amaçlanmıştır.

1.2 Literatürdeki Çalışmalar

Deneysel analiz yöntemlerinin yüksek maliyetli, karmaşık ve çok fazla zaman gerektirmesinden dolayı, son yıllarda biyoinformatik alanında veri analizi alanlarında yapılan çalışmalar avantajlı hale getirmiştir.

Yapıların tahmini için hesaplamalı ve istatistiksel modellemeye yönelik çalışmalar ilk olarak 1973 yılında Krigbaum ve Kuntton tarafından Lineer regresyon analizi ile proteinlerin ikincil yapısının tahmini için kullanılmıştır (Cai, Liu, Xu, & Chou, 2002). Bu probleme yönelik takip eden çalışmada, Chou ve Fasman % 52 başarı ile üçüncül yapı ve ikincil yapı tiplerini frekans tabanlı istatiksel algoritmaları kullanarak elde etmiştir (Chou & Fasman, 1978). Protein yapılarında yapay sinir ağlarına yönelik uygulamalar Stormo ile başlamıştır. Stormo çalışmasında Rosenblatt tarafından 1950 yılının sonlarında geliştirilen algılamalı-öğrenme tabanlı algoritmayı kullanmıştır (Stormo, Schneider, Gold, & Ehrenfeucht, 1982). Genomik informatik alanındaki ilk çalışmalarda çoğunlukla algılamalı ve geri yayılımlı ağların kullanımıyla ilgilenilmiştir (Wu & McLarty, 2012).

1988 yılında, Qian ve Sejnowski yapay sinir ağı modelleri kullanılarak belirlenen homolog olmayan verilerin protein ikincil yapısının (a-helix, ß-sheet ve coil) tahminine yönelik çalışmalar yapmıştır (Qian & Sejnowski, 1988). Çalışmalarında homolog olmayan protein dizisi verilerinde ön işlem olarak pencere kaydırma metodu uygulamışlar ve yapay sinir ağları ile birlikte % 64,3 oranında başarı elde etmişlerdir.

20. yüzyılın sonlarında ikincil yapının tahmini % 65 civarını geçememiştir. Rost ve Sander proteinlerin evrimsel bilgileri kullanarak sınıflandırma performansın da artış elde etmişlerdir. Yaptıkları çalışmada iki katmanlı ileri beslemeli yapay sinir ağlarını kullanmışlar ve 7 kat çapraz doğrulama metodunu uygulayarak % 70,8 doğruluk ile tahmin etmişlerdir (Rost & Sander, 1993).

Bu alandaki gelişmeler devam ederken, araştırmacılar yapay sinir ağları ve istatistiksel metotlar üzerine çalışmalarını genişletmiştir. Krogh ve Riis tarafından protein yapısı tahminine yönelik çalışmalarda yapay sinir ağları ve çoklu dizi hizalama yöntemleri kullanılmıştır. Yaptıkları çalışmalar neticesinde protein yapısının tahmininde % 80 doğruluk elde etmişlerdir (Riis & Krogh, 1996).

1999 yılında Jones ve arkadaşları proteinlerin ikincil yapılarının tahmini için iki katmanlı yapay sinir ağlarını PSM ile eğiterek uygulamışlardır (Jones D. T., 1999). Psi-Blast' dan faydalanarak amino asit konum bilgilerinden PSM görüntü matrisleri elde edilmiştir. Jones tarafından önerilen bu metot PSIPRED olarak isimlendirilmiştir. Üçüncül yapı tahmininde % 76,5 ve % 78,3 seviyesinde bir sonuç elde edilmiştir. Daha sonraki çalışmalarda elde edilen PSM görüntü matrisleri ikincil yapı tahmini için Destek Vektör Makinelerin (DVM) de kullanılmıştır (Ward, McGuffin, Buxton, & Jones, 2003). DVM ile elde edilen eğitim kümesi kullanılarak üçüncül yapı tahmininde % 77 üzerinde bir doğruluk elde etmişlerdir.

Protein yapısının oluşturulmasında amino asit sekans sırasının etkili olduğu bilindiğinden, protein yapısındaki düzensizliklerin belirlenmesinde ayrı bir öznitelik olarak ortaya çıkmıştır. Birçok akademik çalışmada protein düzensizlikleri için hesapsal yöntem önerilmiş ve çoğunlukla dizilim sıralarının önemli olduğu görülmüştür. (Kaya, 2008).

Makine öğrenmesi yöntemi olan yapay sinir ağları, protein yapı tespiti problemi'nin çözümünde genellikle tercih edilen bir yöntemdir (Kaya, 2008). Bu problemin çözümüne yönelik çalışmaların çoğunluğunda düzensiz amino asit yoğunlukları, karmaşıklıkları ve değişimleri ikincil yapı için etkili olduğu bilindiğinden bunun üzerine bilimsel çalışmalar yoğunluk göstermektedir.

Sınıflandırmadaki başarı kalitesini arttırmak için son zamanlarda daha etkili öznitelikler ve kararlı tahmin yapıları oluşturulmaya çalışılmıştır. Örneğin, belirli bir konumu kullanarak pencereleme metoduyla PSM matrislerinden yeni ve düzensizlik tahminini artırıcı özellik vektörleri oluşturulmuştur (Jones & Ward, 2003). Bugüne kadar düzensiz bölgeleri tespit etmek için PONDRs, DisEMBL, GlobPlot, DISOPRED2, FoldIndex, RONN, DisPRO, PreLink, DisPSSMP gibi çok bilinen tahmin araçları geliştirilmiştir.

Protein düzensizliği problemine yönelik ilk çalışma PONDR olarak adlandırılan sınıflandırma aracıdır (Romero, Obradović, Kissinger, Villafranca, & Dunker, 1997). Çalışmada X-Ray karakterize yöntemi ile bölümlenmiş düzensiz protein bölgelerinden elde edilen veri seti kullanılmıştır. Sınıflandırıcı modeli olarak ileri beslemeli yapay sinir ağları eğitilmiş, uygulanan test veri kümesiyle % 58 oranında bir başarı elde etmişlerdir. Tahminde 8 amino asit dizisinin yoğunlukları kullanılarak iki öznitelik elde etmişlerdir.

2003 yılında Max-Delbrück Moleküler Tıp Merkezi tarafından proteindeki düzensizliklerin tespitine yönelik DisEMBL isimli hesaplamalı tahmin aracı geliştirilmiştir. Düzensizliklerde sıralı kısa motifler kullanılarak protein yapısı hakkında önemli bilgiler elde edilmiştir (Linding, ve diğerleri, 2003).

Proteinlerdeki moleküler tanıma tabanlı algoritmaların kullanılması, uzun protein yapılarının tahmininde probleme neden olmuştur. Çünkü bu büyük yapıların tahmini çok uzun sürebilmektedir. DISpro tarafından geliştirilen algoritma ile protein veri bankasından elde edilen veri seti ile protein içerisindeki amino asitlerin konumu ve uzunluklarından istatiksel özellikler çıkarılmıştır. Tek katmanlı özyinelemeli yapay sinir ağları kullanılarak eğitilmiş daha sonra yapılan çapraz doğrulama metodu ile % 92.8 civarında bir doğruluk elde edilmiştir (Cheng, Sweredoski, & Baldi, 2005).

Jaime Prilusky ve arkadaşlarının geliştirdiği FoldIndex tahmin aracında, Uversky algoritması kullanılmıştır. Bu algoritma kullanılarak bir protein üzerindeki hidrofobik özelliğinden faydalanılmış ve protein dizisindeki amino asit dizilimlerinin ağırlık vektörleri çıkarılmıştır. Burada tahmin için geliştirilen yöntemin düzensiz bölgeleri yüksek doğrulukla tahmin etmesini sağlaması için kaydırmalı pencereleme metodu uygulanmıştır. Geliştirilen yöntemin başarı oranları incelendiğinde düzensizliği % 77 oranında düzenli aminoasitleri % 88 oranında tespit ettiği görülmüştür (Prilusky, ve diğerleri, 2005).

Protein yapılarındaki istatiksel analiz metotları, protein düzensiz bölgelerinin tahminini olumlu yönde etkilediğini göstermiştir. Yang ve arkadaşlarının geliştirdiği yapay sinir ağı sınıflandırma metodu kullanılarak örüntü tanıma modeli geliştirilmiştir. RONN olarak isimlendirilen metot, tespit edilen protein dizilerinden sabit uzunluklu bir pencereleme metodu kullanılarak ilişki matrisleri çıkarılmasını sağlamaktadır. Uygulamada protein veri bankasından elde edilen 80 adet protein kullanılmış, 80 protein içerisinden elde edilen öznitelik matrisleri ile eğitilmiştir. Daha sonra test veri seti ile test edilmiştir. Elde edilen performans sonuçları incelendiğinde % 84,9 ve % 78,9 doğrulukla başarı elde edildiği görülmüştür (Yang, Thomson, McNeil, & Esnouf, 2005).

Büyük ölçekli protein verilerinin çok olması verilerin analizinde farklı yollar izlenmesini gerektirmiştir. Bu nedenle veri setindeki uzun dizilimli ve kısa dizilimli veri seti ayrılarak başarı sonuçları karşılaştırılmıştır. IUPForest olarak adlandırılan araç veri setindeki protein dizilerinden kural tabanlı karar ağaçlarını elde edilniştir. Bu model ile elde edilen sonuçlar çapraz doğrulama metodu ile genelleştirilmiştir. Bu yaklaşımla yüksek ölçekteki protein dizilerinde doğruluk oranını arttırılmıştır (Han, Zhang, Norton, & Feng, 2006).

Protein verilerindeki dizilimlerin birbirleri ile ilişkili olduğu bilinmektedir. Bu proteinlerin düzensizliklerin tespitinde önemli rol oynadığı görülmüştür. Bu nedenle geliştirilen yöntemle elde edilen öznitelikler Spritz aracında DVM ile sınıflandırılmış ve CASP6 veri seti ile test edilerek düzensiz bölgeleri ortalama % 99, düzenli bölgeleri ise % 98 oranında tespit ettiği görülmüştür (Vullo, Bortolami, Pollastri, & Tosatto, 2006).

Deneysel çalışmalar neticesinde farklı fizyolojik koşullar altında değişim gösteren düzensiz protein bölgeleri olduğu görülmüştür. Proteinler, hücre sinyal iletimi, transkripsiyonel ve translasyonel düzenleme gibi çeşitli biyolojik süreçlerin de tamamında etkili oynamaktadır. 2007 yılında POODLE-L olarak adlandırılan makine öğrenmesi tabanlı düzenli ve düzensiz bölge tahmin aracı geliştirilmiştir.10 farklı fizikokimyasal özellik kullanılmış ve 2 seviyeli DVM sınflandırıcısı da eğitilmiştir. Uzun düzensiz bölgelerin sınıflandırma performansı Matthew’s korelasyon katsayısı kriterine göre 0,658 olarak elde edilmiştir (Hirose, Shimizu, Kanai, Kuroda, & Noguchi, 2007).

PrDOS olarak adlandırılan araç amino asit dizilerinden elde edilen protein düzensiz bölgelerini iki farklı sınıflandırıcı kullanarak tahmin eden bir sistem geliştirmiştir (Ishida & Kinoshita, 2007)

OnD-CRF olarak isimlendirilen düzensiz bölgelerin tahminine yönelik geliştirilen uygulamada şartlı rastgele alan metodu kullanılarak eğitim yapılmıştır. Eğitim kümesinde amino asit dizilerinin sırası, konumu gibi birçok özellik öznitelik vektöründe kullanılmıştır. Uygulamada başarım sonucunu diğer sınıflandırıcılar ile karşılaştırmak için CASP7 veri seti ile test edilmiştir.(Wang & Sauer, 2008).

2009 yılında Deng ve arkadaşları tarafından ab-initio dizi tabanlı bir tahmin metodu geliştirilmiştir. Sınıflandırma metodu PreDisorder olarak adlandırılmıştır. CASP8 protein verisi üzerinde test edilerek test edilmiştir (Deng, Eickholt, & Cheng, 2009).

Proteindeki düzensiz bölgelerin tahmini için pek çok deneysel metot kullanılmıştır. Bununla birlikte düzensizliklerin tahmini için protein araştırmaları da hız kazanmıştır. PONDRFIT olarak adlandırılan çalışmada sınıflandırıcıların sonuçları birleştirerek bir sınıflandırma algoritması geliştirilmiştir (Xue, Dunbrack, Williams, Dunker, & Uversky, 2010).

MoRFpred olarak adlandırılan tahmin aracında AIndex, PSM matrisleri kullanılmıştır. Elde edilen 24 özellik vektörü DVM sınıflandırıcısında eğitilmiştir. Daha sonra test veri setleri uygulandığında % 94,6 ve % 88,9 başarı elde edilmiştir (Disfani, ve diğerleri, 2012).

Düzensiz bölgelerin tahminine yönelik kısa ve uzun proteinler için ayrı ayrı sonuçlar elde edilmesi problemin genellenmesi açısından önemli olmuştur. SPINED olarak isimlendirilen tahmin aracında yapay sinir ağları sınıflandırıcısı kullanılmıştır. Bu uygulamada SL329 veri seti ile yapılan test işlemi sonucunda %88,6 başarı elde edilmiştir (Zhang, ve diğerleri, 2012).

DNdisorder tahmin aracında protein düzensizliklerine yönelik yeni bir dizi geliştirilmiştir. Düzensiz bölgelerin tahmini için makine öğrenmesi yöntemi olarak yapay sinir ağlarını kullanılmıştır. 723 protein üzerinde yapılan çapraz doğrulama metodu ile % 82’ lik bir doğruluk elde edilmiştir. Geliştirilen yöntemin doğruluğunu tespit etmek için CASP 9 ve CASP10 veri setleri kullanılarak test edilmiştir (Eickholt & Cheng, 2013).

Ökaryotik proteinlerin büyük bir çoğunluğunda düzensiz bölgeler bulunmaktadır. Bu protein veri seti DISOPRED3 olarak adlandırılan çalışmada kullanılmıştır. 20 amino asitten çok içerisinde düzensiz bölge içeren protein yapıları seçilerek eğitim kümesinde kullanılması amaçlanmıştır. Sınıflandırıcı olarak DVM tabanlı bir yapı kullanılmıştır. Oluşturulan araç CASP9 ve CASP10 veri seti kullanılarak test edilmiştir (Jones & Cozzetto, 2015).

1.3 Hipotez

Proteindeki düzensiz bölgelerin bilgisayar ortamında tahmini probleminde yeni öznitelik kodlama yöntemleri geliştirilmesi problemin yüksek doğrulukla tespiti açısından kritik öneme sahiptir. Bu tez çalışmasında yeni öznitelik kodlama yöntemleri geliştirilmesinde PSM, aminoasitlerin fizikokimyasal ve SPINE-X protein özelliklerinin kullanılabileceği öngörülmektedir. Ayrıca, protein dizileri sonlu bir sinyal olarak düşünüldüğünde Fourier dönüşümlerinden olan dalgacık teoreminin de öznitelik vektörü oluşturmada kullanılabileceği düşünülmektedir.

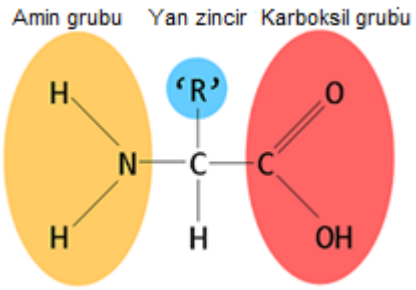
2. PROTEİNLERDEKİ DÜZENSİZ BÖLGELER

Biyolojik diziler, nükleotid dizileri ve amino asit dizilerini kapsar. Nükleotid dizileri DNA moleküllerini, amino asit dizileri ise protein moleküllerini meydana getirir. Canlıların neredeyse tüm biyolojik etkinliklerini belirleyen proteinler, DNA moleküllerinin yorumlanması ile oluşturulurlar. Canlıların da görünümleri, hareket tarzları, beslenme alışkanlıkları ve çoğalma biçimlerindeki çeşitlilik bu farklılıkların sonucudur: Bir canlıyı tanımlayan ve onu diğer canlılardan farklı yapan bilginin neredeyse tamamı doğrudan hücrelerinde veya dolaylı olarak da sentezledikleri proteinlerde saklıdır. Bu nedenle proteinlerdeki düzensiz bölgeler vücut yapımızda önemli bir etken olmaktadır. Bu sebeple bu yapıları iyi tanımamız ve yapıları haklarında detaylı bilgie dinmemiz gereklidir. Biyolojik verilerin iyi bir şekilde anlaşılması, nükleotid dizileri ya da karşılık geldikleri amino asit dizilerinin, yapısal ve işlevsel ilişkilerinin ortaya çıkarmasına olanak sağlar (Mount, 2001).

2.1 Aminoasitler

Proteinler çok sayıda amino asit olarak adlandırılan küçük yapı taşlarından oluşur. Doğada 300'den fazla amino asit vardır. Fakat insanlarda bunlardan yalnız 20 tane bulunmaktadır. Proteinlerde 20 çeşit amino asit bulunabilir.

Bir nitrojen (N) ve iki hidrojen (H) atomu amino grubunu (-NH2) oluşturur. Karboksil grubu (-COOH) ise asit varlığını oluşturur. Amino aside belirleyen yan zincir ise R-grubudur.



Şekil 2.1: Genel amino asit yapısı.

Amino asitler, bir amino asidin karboksil gurubu ile diğer amino asidin amino grubunun reaksiyona girmesi sonucu bir molekül su (H2O) açığa çıkararak, oluşan peptit bağıyla birbirine bağlanır. Peptitler 2 veya daha fazla amino asitten oluşan bileşiklerdir. Oligopeptitlerde 10 veya daha az amino asit bulunur. Polipeptitler ve proteinler 10 veya daha fazla amino asit zincirlerinden oluşmuşlardır. 50'den fazla amino asitten oluşan peptitler, protein olarak sınıflandırılır.

2.1.1 Amino Asitlerin Fizikokimyasal Özellikleri

Bu tez çalışmasında aminoasitlere ait 5 adet fizikokimyasal özellik kullanılmıştır. Bunlar; sterik özelliği, hidrofobiklik yoğunluğu, hacim, kutuplanabilme ve izoelektrik noktasıdır (Meiler, Müller, Zeidler, & Schmäschke, 2001). Bu 5 özellik kullanılarak iki istatiksel parametre (Helix ve Sheet olasılığı) elde edilmiştir. (Fauchere, Charton, Kier, Verloop, & Pliska, 1988). Çizelge 2.1'de kullanılan aminoasitlerin özellikleri verilmiştir.

Çizelge 2.1: Aminoasitlerin 7 farklı özelliğinin sayısal değerleri.

İsim	Ξ^a	α^b	$\upsilon_v{}^c$	π^d	I^e	α^f	β^g
ALA	1,28	0,05	1,00	0,31	6,11	0,42	0,23
GLY	0,00	0,00	0,00	0,00	6,07	0,13	0,15
VAL	3,67	0,14	3,00	1,22	6,02	0,27	0,49
LEU	2,59	0,19	4,00	1,70	6,04	0,39	0,31
ILE	4,19	0,19	4,00	1,80	6,04	0,30	0,45
PHE	2,94	0,29	5,89	1,79	5,67	0,30	0,38
TYR	2,94	0,30	6,47	0,96	5,66	0,25	0,41
TRP	3,21	0,41	8,08	2,25	5,94	0,32	0,42
THR	3,03	0,11	2,60	0,26	5,60	0,21	0,36
SER	1,31	0,06	1,60	-0,04	5,70	0,20	0,28
ARG	2,34	0,29	6,13	-1,01	10,74	0,36	0,25
LYS	1,89	0,22	4,77	-0,99	9,99	0,32	0,27
HIS	2,99	0,23	4,66	0,13	7,69	0,27	0,30
ASP	1,60	0,11	2,78	-0,77	2,95	0,25	0,20
GLU	1,56	0,15	3,78	-0,64	3,09	0,42	0,21
ASN	1,60	0,13	2,95	-0,60	6,52	0,21	0,22
GLN	1,56	0,18	3,95	-0,22	5,65	0,36	0,25
MET	2,35	0,22	4,43	1,23	5,71	0,38	0,32
PRO	2,67	0,00	2,72	0,72	6,80	0,13	0,34
CYS	1,77	0,13	2,43	1,54	6,35	0,17	0,41

a: Sterik parametresi (grafik şekil indeksi)

b: Kutuplanabilirlik

c: Hacmi (Normalize van der Waals hacmi)

d: Hidrofobik etkisi

e: İzoelektrik nokta

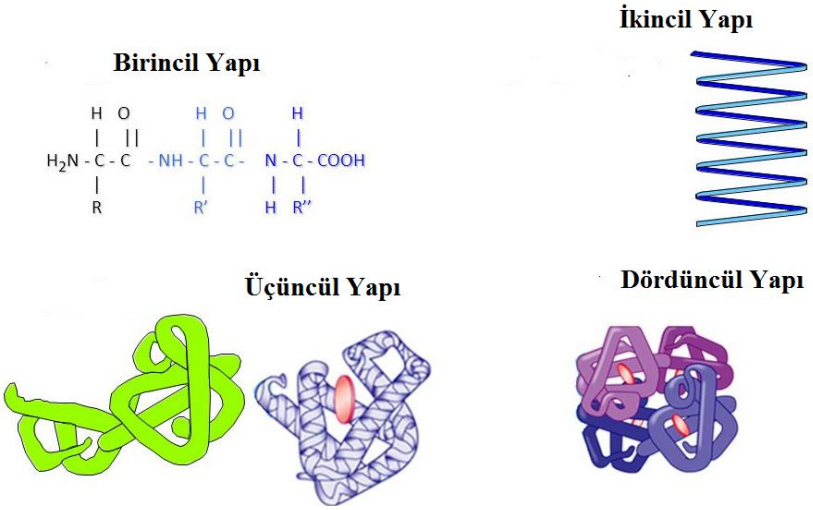
f: Helix olasılığı

g: Sheet olasılığı

2.2 Proteinler

Proteinler, birbirlerine bir zincir şeklinde peptit bağı kullanılarak bağlanmış amino asitlerden oluşan büyük organik bileşiklerdir. Karbon, hidrojen, oksijen ve azottan oluşmaktadırlar. Hücrelerde protein sentezi sonrasında üretilen amino asitlerin birbirine bağlanarak oluşturdukları düz zincir, daha sonra amino asitler arasındaki kimyasal bağların etkisi neticesinde katlanarak protein oluşturmaktadır Proteinlerin bazıları heliks/sarmal yapıda olabileceği gibi küresel yapıda veya antikorlar gibi Y şeklinde de olabilirler. Proteinler üç boyutlu yapılarındaki girinti çıkıntıları kullanarak ya başka proteinlerle ya da alıcı moleküler yapılara bağlanarak hücre içi faaliyetlerini sağlarlar. Anahtar kilit ilişkisine benzeyen yapılarla proteinlerin birbirlerine ya da diğer moleküllere bağlanıp ayrılması, proteinlerin üç boyutlu yapılarını çok önemli olmasını sağlar. Bir proteinin aktif bölgesindeki sadece bir amino asidin bile yerinin değişmiş olması, proteinin şeklini değiştirip tamamen temel görevini yapmasını engellemektedir. Bu nedenle protein sentezi sonrası zincir gibi olan amino asit dizisinin katlanarak olması gereken şeklini alması çok önemlidir.

Proteinlerin 4 farklı çeşit yapısı vardır. Bunlar birincil, ikincil, üçüncül ve dördüncül yapılardır.



Şekil 2.2: Proteinlerin birincil, ikincil, üçüncül ve dördüncül yapıları.

2.2.1 Proteinlerin Düzensiz Bölgeleri

Proteinlerin insanın temel yapısını oluşturması yapısının iyi bir şekilde tanınması ve anlaşılmasını gerektirmiştir. Bu nedenle düzensiz bölgelerin tahmini ciddi bir problem olarak önümüzde durmaktadır.

Bu bölgeler de, dizilimlerin karmaşık olması özellikle fiziko-kimyasal özelliklere sahip olma eğilimi göstermesi ve amino asit oluşumunu olumsuz yönde etkiler.

Üçüncül yapıların tahmininde düzenli ya da düzensiz bölgeler göz ardı edilmiştir (Romero, ve diğerleri, 2001). Üç boyutlu yapısı bulunmayan bir protein doğal olarak katlanmamış ya da kendiliğinden düzensiz olarak anılmıştır (Dunker, Romero, Obradovic, Garner, & Brown, 2000). Amino asit dizilerindeki bu tür bozukluklar biyolojik fonksiyonlardaki değişimlere de neden olmuştur. Bu tür nedenlerin varlığı da "Proteindeki düzensiz bölgeler" problemini ortaya çıkarmıştır (Li, Brown, Obradovic, Garner, & Dunker, 2000).

Düzensiz bölgeler, çeşitli araştırmalar sonucunda bazı hastalıklarla ilişkilendirilmiş ve yapılan araştırmalar sonucunda, düzensiz proteinlerin kardiyovasküler hastalıklar, kanser, diyabette Miletus, nörodejeneratif hastalıklar ve otoimmun rahatsızlıklarla ilişkisi tespit edilmiştir (İmer & Çavas, 2009).

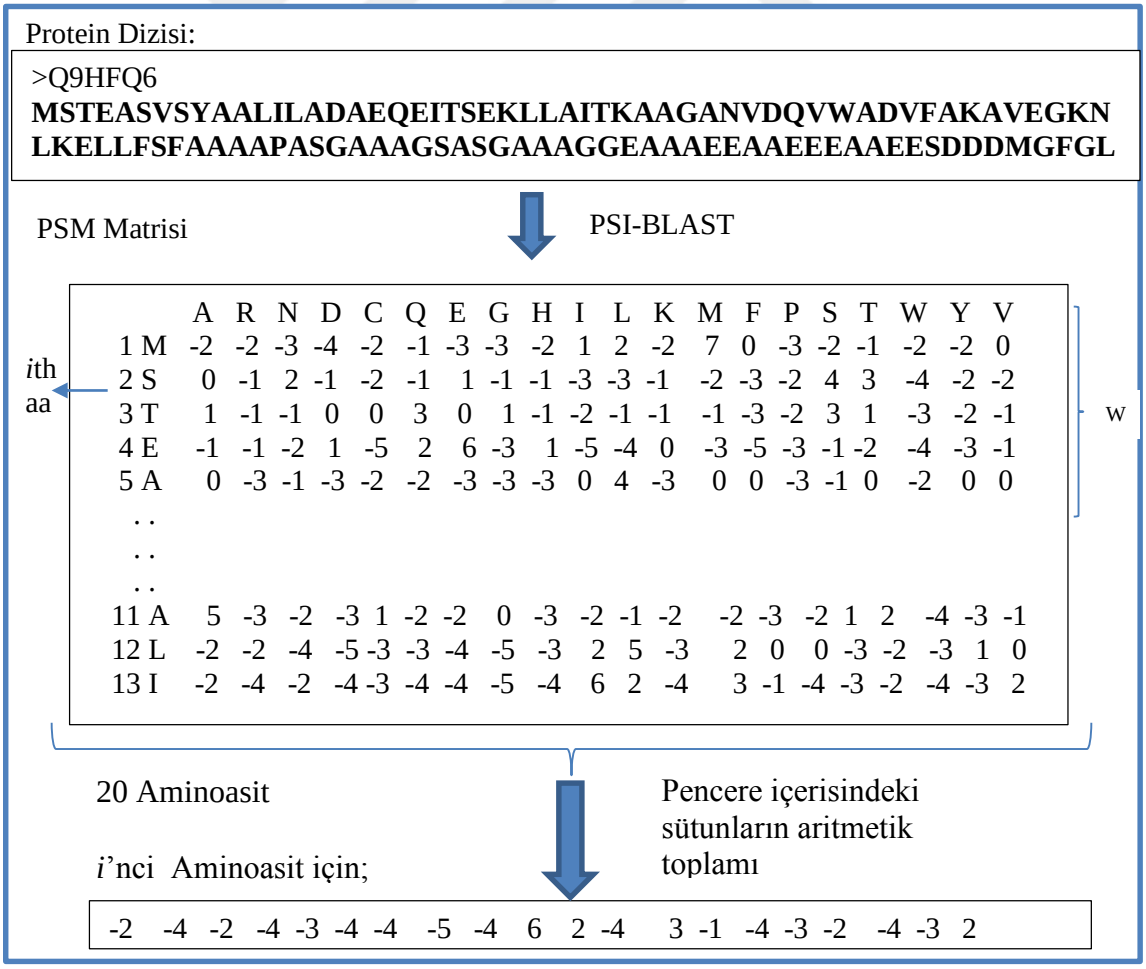
2.3 PSM Matrisi

Gelişim tabanlı profillerin kullanılması düzensiz bölgelerin tahminindeki başarı oranını arttırdığı görülmüştür (Jones & Ward, 2003). Bu nedenle gelişim tabanlı bir profil olan PSM matrisi kullanılmıştır.

PSM ile belirli bir konumda kalıntı korunması için bir ölçü oluşturmaktadır ve 20 boyutlu vektör ile temsil edilmektedir (Altschul, ve diğerleri, 1997). Buradaki 20 boyutlu vektör 20 farklı amino asidin mutasyona karşı korunması için temsil olasılıklarını verir. Bu nedenle *N* uzunluğundaki bir aminoasit dizisinin PSM matrisi oluşturulduktan sonra *Nx20* görünümünde bir matris oluşmaktadır.

Dizilim içerisinde *i'*nci konumdaki aminoasit için, X_{ia} öznitelik değeri eşitlik 2.1 ile hesap edilmiştir.

$$X_{ia} = \frac{1}{w}\sum_{j=w_s}^{w_f} M_{ja} \qquad \textbf{(2.1)}$$



Şekil 2.3: Pencereleme metoduyla Protein 'den PSM profilinin çıkarılması.

2.4 SPINE-X Aracı

SPINE-X, protein ikincil yapı tahminine yönelik sparks-lab tarafından geliştirilen bir araçtır (Faraggi, Yang, Zhang, & Zhou, 2009) (Zhang, Faraggi, & Zhou, 2010). Bu araç kullanılarak protein dizilerindeki her amino asit için 15 yeni öznitelik vektörü oluşturulmuştur.

Oluşturulan öznitelik vektörlerin de; İkincil yapı özellikleri olan, solvent erişilebilirlik özelliği, omurga torsiyon açısı, rafine torsiyon açıları, maksimum entropy tahmini, torsiyon açısı dalgalanmaları, solvent erişe bilirlik özelliği standart sapması, omurga torsiyon açısı standart sapması, torsiyon açısı dalgalanma standart sapma değerleri alınmıştır.

2.5 Veri Setleri

Protein veri bankası tarafından düzensiz bölge olması mümkün olan protein bölgeleri X-Ray Crystallography yöntemi ile tespit edilebilmektedir. Ancak bu bölgelerin kesin sonuçlarla tespit edilmesi mümkün olmamaktadır. Bu nedenle elde edilen veri seti literatürde bu problem için kullanılan ve kabul gören bir çalışmadan alınmıştır.

Bu çalışmada eğitim kümesi için SL Benchmark veri seti indirilmiştir (Sirota, ve diğerleri, 2010). Disprot tarafından düzensiz bölgeleri düzenlenerek SL veri seti elde edilmiştir. SL veri setini filtrelemek için blastclust aracı kullanılmıştır. 329 adet zincirden oluşun SL329 veri seti elde edilmiştir. Her bir protein arasında %25 oranında bir özdeşlik vardır. SL329 veri seti özellikleri Çizelge 2.2 de verilmiştir.

Çizelge 2.2: SL329 veri seti özellikleri.

SL329	
Protein Zincir Sayısı	329
Düzenli Amino Asit Sayısı	51292
Düzensiz Amino Asit Sayısı	39544
Toplam Amino Asit Sayısı	90836

Elde edilen SL329 veri seti içinde protein dizilimlerinin birçoğunda düzenli ve düzensiz aminoasit sayısı oranları çok fazla değişim gösterdiği görülmüştür. Bu nedenle bu çalışmada SL329 veri setindeki her bir protein kendi içerisinde değerlendirilerek yeniden alt veri seti oluşturulmuştur. Buradaki amaç, protein dizileri içinden %40 ve %60 arasındaki düzensizlik oranına sahip proteinler seçerek yeni bir eğitim kümesi oluşturmaktadır. Oluşturulan yeni veri seti SDR38 olarak isimlendirilmiştir.

Çizelge 2.3: SDR38 veri seti özellikleri.

SDR38(SubDisorderRate-SDR)	
Protein Zincir Sayısı	38
Düzenli Amino Asit Sayısı	3400
Düzensiz Amino Asit Sayısı	6712
Toplam Amino Asit Sayısı	10112

Aynı şekilde SL329 veri seti içerindeki düzensiz protein oranı % 20 ve % 80 aralığında olanlar seçilerek yeni bir veri kümesi oluşturulmuştur ve SDR147 olarak isimlendirilmiştir.

Çizelge 2.4: SDR147 veri seti özellikleri.

SDR147	
Protein Zinciri Sayısı	147
Düzenli Amino Asit Sayısı	19274
Düzensiz Amino Asit Sayısı	22984
Toplam Amino Asit Sayısı	42258

Bu çalışmada ayrıca Protein Structuer Prediction Center tarafından sunulan CASP10 veri seti kullanılarak önerdiğimiz yöntem test edilmiştir. Bu problemin çözümüne yönelik yapılan çalışmalarda, CASP10 veri seti ile test edilerek sunulmuş olması çalışmalardaki performans sonuçlarının karşılaştırılması açısından önemlidir. Veri setinin özellikleri Çizelge 2.5'te sunulmuştur

Çizelge 2.5: CASP10 veri seti özellikleri.

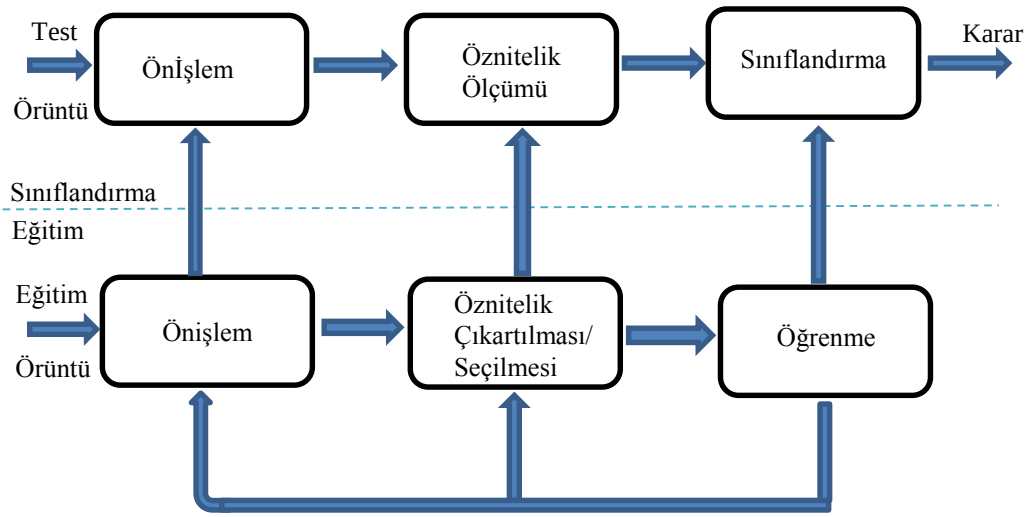
CASP10	
Protein Zinciri Sayısı	95
Düzenli Amino Asit Sayısı	22673
Düzensiz Amino Asit Sayısı	1597
Toplam Amino Asit Sayısı	24270

3. ÖZNİTELİK KODLAMA

Örüntü tanıma, nesneleri farklı sınıflandırma amacına dayalı bir bilim dalıdır. Sınıflandırılmak istenen nesneler biyomedikal verilerden, elektromanyetik işaret dalgalarına kadar çeşitlilik gösterebilir. Örüntü tanıma, mühendislik araştırma, geliştirme konularında en uygun kararın verilmesinde önemli bir rol oynamaktadır.

Çalışma konusu olan proteinlerin düzensiz bölgelerinin tespitine ilişkin problemin çözümünde bilgisayar destekli olarak çalışılmıştır. Şekil 3.1'de görüldüğü gibi örüntü tanıma sistemi eğitim ve sınıflandırma olmak üzere iki kısımdan oluşmaktadır. Önişlem kısmının amacı, örüntünün en iyi özelliklerini belirleyerek, anlaşılır ve işlenebilir hale getirmektir. Bu amaç doğrultusunda protein verileri üzerine araştırma yapılıyorsa örüntü verisi normalize edilerek bir kodlamaya tabi tutulabilir (Öznitelik kodlama yöntemi) veya elektronikte işaret işleme teknikleri üzerine araştırma yapılıyorsa işaretler üzerinde gürültü filtreleme işlemleri uygulanabilir. Bütün bu işlemlerde, örüntünün uygun bir şekilde temsil edilmesi amaçlanmaktadır (Duda, Hart, & Stork, 2001).

Bir öğrenme sisteminin modellemesi, sınıflandırma veya bağlanım (regression) problemi olabilmektedir (Wlodawer & Erickson, 1993). Bağlanım analizi, iki veya daha fazla değişken arasındaki ilişkileri ölçmek için kullanılan istatistiki bir yöntemdir. Sınıflandırma ise örüntü tanıma sürecine dâhil edilen giriş verilerinin, süreç sonunda, tanımlanmış olan sınıflardan hangisine ait olduğunun tahmin edilmesi sürecini kapsayan istatistiki bir yöntemdir.



Şekil 3.1: Örüntü tanıma sistemi.

Öznitelik, örüntüye ait ölçülebilir veya gözlenebilir veriler olarak tanımlanmaktadır. Eğitim sürecinin, öznitelik çıkartılması/seçilmesi kısmında temsil edilen örüntü verileri için en uygun öznitelikler tespit edilir ve sınıflandırıcı öznitelik uzayını bu yönde çeşitli bölümlere ayırır. Test kısmında, eğitilmiş sınıflandırıcı, giriş örüntülerini öznitelik ölçümlerine göre hangi sınıflara ait olduğuna karar verir.

3.1 Dalgacık Teoremi

Dalgacık teorisinin kökeni bir Fransız matematikçi olan Jean Baptiste Joseph Fourier (1768-1830) tarafından geliştirilen harmonik analiz yöntemiyle başlamıştır (Gargour, Gabrea, Ramachandran, & Lina, 2009). Kosinüs ve sinüs fonksiyonlarının ağırlıklı toplam açısından herhangi bir periyodik fonksiyonu tanımlayan bir yöntem geliştirmiştir. Daha sonra 1909 yılında Alfred Haar Haar Dalgacık ailesini geliştirmiştir (Haar, 1910). Bunun sonucunda Haar dalgacık ailesi ölçeklendirme özelliğinin modelleme fonksiyonlarında daha doğru sonuç verdiği kanıtlamıştır.

Dalgacık teoreminin gelişim kronolojisi incelendiğinde temelini fourier dönüşümü oluşturmaktadır. Fourier dönüşümü bilindiği gibi gerçek ölçülen sinyaller zaman ya da frekans domenin de değişimler incelenmektedir. Bu sinyaller durağan olmayan sinyallerdir. Fourier dönüşümü bu tarz sinyalleri frekans içeriklerini ya da frekans bileşimlerini bulmak için kullanılmaktadır. Fourier dönüşümü Eşitlik 3.1 kullanılarak hesaplanmaktadır. *F* frekans ve birimi Hertz dir. Ωt radyan cinsinden değeridir.

$$FT\{x(t)\} = X(\Omega) = \int_{-\infty}^{+\infty} x(t)e^{-j\Omega t}dt, \;\; \Omega = 2\pi F \tag{3.1}$$

Fourier dönüşümü, belirli frekansı bulunan zaman aralıklarında uygun değildir. Bu nedenle dalgacık dönüşümü fourier dönüşümünün çözünürlük problemini aşmak için alternatif yaklaşım önermiştir.

$$CWT_x^{\varphi}(a,b) = X_{\varphi}(a,b)$$
$$= \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t) \left[\varphi^* \left(\frac{t-b}{a}\right)\right] dt = \langle x(t), \varphi_{a,b}^*(t) \rangle \tag{3.2}$$

a ve b sırasıyla ölçekleme ve çeviri parametreleridir. $\varphi_{a,b}^*(t)$ ve $\frac{1}{\sqrt{a}}\varphi^*\left(\frac{t-b}{a}\right)$ ana dalgacık yani temel fonksiyondur. Diğer pencere fonksiyonlarını üretmek için bir protatiptir. Tüm pencereler, genişlemiş sıkıştırılmış ve ana dalgacık sürümlerinden türetilmektedir.

Özetle dalgacık analiz teknikleri geleneksel fourier dönüşümüne göre daha iyi bir performans göstermektedir (Sifuzzaman, Islam, & Ali, 2009). Dalgacık analizi ile fourier dönüşümünün karşılaştırılması Çizelge 3.1 de gösterilmiştir.

Çizelge 3.1: Fourier dönüşümünün ile dalgacık analizinin karşılaştırılması.

Özellikler	Fourier Dönüşümü	Dalgacık Analizi
Durağan Sinyaller	Evet	Evet
Durağan Olmayan Sinyaller	Hayır	Evet
Zaman Domeni	Hayır	Evet
Frekans Domeni	Evet	Evet
Ölçek	Evet	Evet
Kaydırmalı	Hayır	Evet

3.2 Sınıflandırma Algoritmaları

Makine öğrenmesi bilgisayar bilimlerinin bir alt alanıdır ve veriler üzerinde tahmin yapabilen algoritmaların yapılarını inceler (Kohavi & Provost, 1998). Arthur Samuel makine öğrenmesini, "Bilgisayarlara açıkça programlanmadan öğrenme yeteneği katan bir çalışma alanıdır." biçimin de tanımlamıştır (Simon, 2013).

Bilgisayarlı tahmin ve hesaplamalı istatistik birbirleriyle oldukça ilişkilidir ve hatta çoğu durumda bazı durumlarda örtüşür. İstenmeyen posta filtrelemesi, optik karakter tanıma, arama motorları, bilgisayarlı görme, örüntü tanıma, bilgisayarlı hastalık teşhisi gibi birçok alanda uygulaması mevcuttur (Wernick, Yang, Brankov, Yourganov, & Strother, 2010).

Sınıflandırma algoritmaları temel olarak iki ana dalda incelenir. Bunlar;

- Danışmanlı sınıflandırma
- Danışmansız sınıflandırma

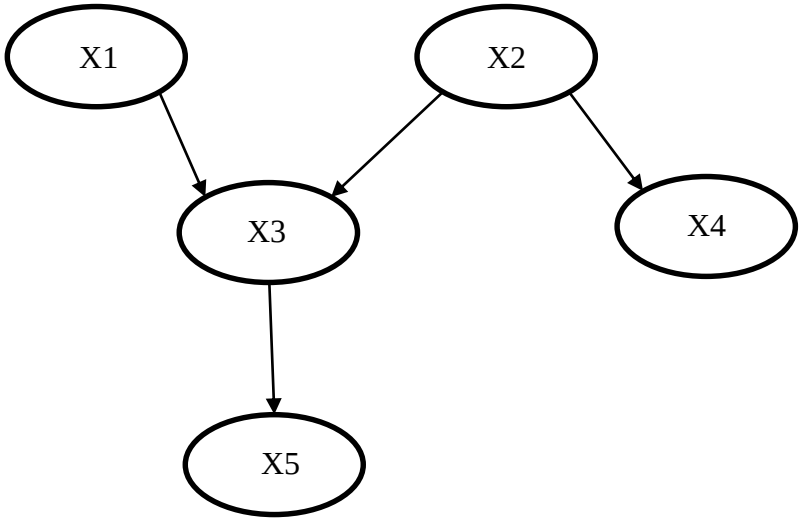
Danışmanlı sınıflandırmada girdiler için hedef çıktılara karşılık bir sonuç üretilir. Danışmansız algoritmalarda ise amaç; herhangi bir sınıf verisine ait olmayan girdilerden, benzer doğal özellik ya da mesafeleri içerenleri bir araya getirmektir. Kümeleme, danışmansız sınıflandırma tekniğine ait bir örnektir (Alpaydin, 2014). Bu çalışmada verilerin sınıf bilgisine sahip olmasından ötürü danışmanlı sınıflandırma teknikleri kullanılmıştır.

Proteinlerin düzensiz bölgelerinin tahmini için geliştirilen öznitelik kodlama yöntemleri Bayes ağları, Naive Bayes, DVM, k-EYK ve sıralı minimal optimizasyom sınıflandırıcı algoritmaları ile sınanmıştır. Böylece her bir sınıflandırıcının geliştirilen öznitelik kodlama yöntemlerine göre başarımları elde edilmiştir.

3.2.1 Bayes Ağları

Bayes Ağı, düğümler aracılığıyla değişkenlerin istatiksel gösteriminin yapıldığı grafiksel kısım ve değişkenlere ait koşullu olasılık tablolarından oluşmaktadır. Bayes ağlarında grafiksel kısım ağın yapısını oluşturmaktadır. Ağda iki düğüm birbirine bağlantı yardımıyla bağlandığında, bağlantının başlangıcı bulunduğu nokta ana düğüm, bağlantının bitişinde bulunan nokta ise alt düğüm olarak adlandırılır.

Şekil 3.2'de X1, X2, X3, X4 ve X5 değişkenlerinden oluşan örnek bir Bayes ağının gösterimi yapılmaktadır.



Şekil 3.2: X1, X2, X3, X4 ve X5 değişkenlerinden oluşan örnek Bayes Ağı yapısı.

Bu ağ içerisinde X1 ve X2 değişkenleri X3 değişkeni oluşturmakta, X3 değişkeni ise X5 değişkenini oluşturmaktadır. Ayrıca X4 değişkeninde X2 değişkeninin alt değişkenidir.

Şekilde değişkenlerin sahip oldukları koşullu olasılık dağılımları, P(X1), P(X2), P(X3| X1, X2), P(X4| X2) ve P(X5| X3) oluşturulmalıdır. Ağda yer alan bir değişkenin, başka bir değişkenle arasında herhangi bir ilişkinin bulunmaması o değişkenin ağda yer alan diğer değişkenlerle arasında istatiksel bir ilişki bulunmadığını, dolayısıyla ağda gerçek olasılık dağılımı ile yer aldığını gösterir (N. Friedman, 1997).

Grafikteki bu koşullu bağımlılıklar genellikle bilinen istatistiksel ve hesaplama yöntemleriyle belirlenir. Bu nedenle Bayes ağları çizge kuramının, olasılık teorisinin, bilgisayar bilimlerinin ve istatistiğin temel ilkelerini bir arada kullanmaktadır (Ben-Gal, 2007). Bayes ağları Eşitlik 3.3 ile hesaplanır;

$$P(U) = \prod_{u \in U} p(u|pa(u)) \qquad \textbf{(3.3)}$$

Burada $BP = \{p(u|pa(u))|u \in U\}$ bağıntısına ait olasılık çizelgesini içeren bir ağdır. Ayrıca, $pa(u)$, U kümesinin bir kısmını içeren alt kümesidir.

3.2.2 Naïve Bayes

Naïve Bayes sınıflandırıcısı, Bayes ağlarının basitleştirilmiş bir formudur. Tüm öznitelikler verilen sınıf verilerinden bağımsızdır. Bu yönteme koşullu bağımsızlık adı verilir (Zhang H. , 2005). Diğer danışmanlı makine öğrenmesi yöntemleriyle kıyaslandığında Naïve Bayes sınıflandırıcısı sınıflı eğitim verileri için en basit basit ve güçlü tahmin tekniğine sahip olan sınıflandırıcılardan biridir. Güçlü tahmin yeteneğinin yanında eğitim setinde öznitelik ve sınıflar arasındaki ilişki hakkında iyi bilgiler verebilir (Možina, Demšar, Kattan, & Zupan, 2004). Naïve Bayes fonksiyonu Eşitlik 3.4 ile hesaplanır.

$$f_{nb}(E) = \frac{p(C=+)}{p(C=-)} \prod_{i=1}^{n} \frac{p(x_i|C=+)}{p(x_i|C=-)} \qquad \textbf{(3.4)}$$

Eşitlikte;

E : X öznitelik kümesinin i indisli öznitelik vektöründeki $(x_1, x_2, ..., x_n)$ değerlerin bir kısmından oluşan veri kümesini,

C : Sınıflandırıcı değerleri. $+$ Pozitif sınıfları, $-$ negatif sınıfları,

$f_{nb}(E)$: E giriş değerleri için Naïve Bayes fonksiyonunu temsil etmektedir.

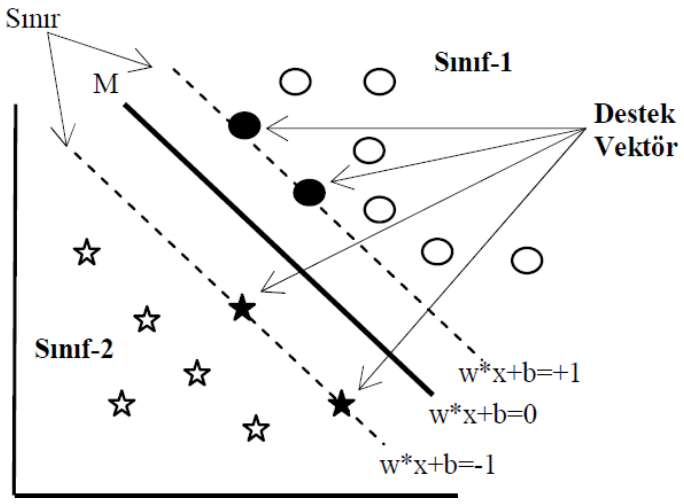
3.2.3 Destek Vektör Makinaları

Destek vektör makineleri (DVM), Vapnik tarafından geliştirilen ve istatistiksel öğrenme teorisine dayanan bir yöntemdir. DVM'nin diğer sınıflandırıcılardan farkı yanlış sınıflandırma ihtimalini en aza indirgeyecek çözümler bulmaya çalışır. DVM yöntemi son yıllarda birçok alanda kullanılmaya başlanmıştır. Resim ve nesne tanıma, ses tanımlama, parmak izi tanıma, el yazısı tanıması ve birçok alanda kullanılmıştır.

Farklı 2 örnek verilmiş sınıftan oluşan veri vektörü n-boyutlu uzayda $\{x_i\} \in R^n$ ve $(x_i, y_i), i = 1,2,3, \ldots\ldots, N$ olmak üzere buna karşılık gelen sınıf etiketi $y_i \in \{+1, -1\}$ olsun.

Bir hiperdüzlem $w^t x + b = 0$ tarafından ayrıştırılabilir. Buradaki w n boyutlu bir vektör, b ise eşik değeridir.

$$y_i(w^t x + b) \geq 1, \qquad i = 1,2,3, \ldots\ldots, N \qquad \textbf{(3.5)}$$



Şekil 3.3: DVM algoritması hiper düzlemi.

Maksimum aralık $2/\|w\|$ alınır. Böylece $\|w\|^2$ minimize ederek maksimum marjin veren hiperdüzlemi bulur. Bu problem Lagrange optimizasyon yöntemi kullanarak ikili sorun çözülebilir

$$Maksimize\ Q(a) = \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i,j=1}^{N} a_i a_j y_i y_j K(x_i, x_j) \qquad \textbf{(3.6)}$$

$K(x_i, x_j)$, kernel fonksiyonudur. Bu denklemde a_i Lagrange çarpanlarıdır. Eğitim setindeki her bir örnek a_i' ye karşılık gelmektedir.

Optimizasyon teorisinin Kuhn-Trucker teorisine göre sıfır olmayan Lagrange çarpanlarına sahip örnekler destek vektör olarak adlandırılır. Eğer a_i en iyi çözüm ise, sınıflandırıcı karar fonksiyonu Eşitlik 3.5 ile ifade edilebilir (Jiang, 2005).

$$\operatorname{sgn}\left(w^T \phi(x) + b\right) = \operatorname{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b\right) \qquad \textbf{(3.7)}$$

3.2.4 Sıralı Minimal Optimizasyon

Sıralı minimal optimizasyon, DVM eğitimleri süresince meydana gelen kuadratik programlama problemlerini çözmek için John Platt tarafından 1998 yılında ortaya çıkarılmış bir algoritmadır (Platt, 1998).

DVM makine öğrenmesi yöntemlerini kullananlar tarafından oldukça ilgi görmüş, önceleri kullanılan yöntemlerden daha az karmaşık ve maliyetli olduğu tespit edilmiştir (Zanni, Serafini, & Zanghirati, 2006) (Rifkin, 2002). Sıralı minimal optimizasyon algoritması Eşitlik 3.8 ile hesaplanır;

Çok terimli kernel kullanarak destek vektör sınıflandırıcıyı eğitmek için SMO Algoritmasını uygular. Bu uygulama global olarak bütün kayıp değerleri yenisiyle değiştirir ve nominal öznitelikleri ikili olanlara dönüştürür. Ayrıca bütün özellikleri önceden belirlenmiş değerlerle normalize eder.

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j K(x_i, x_j) \alpha_i \alpha_j \quad \textbf{(3.8)}$$

Eşitlikte;

α_i : Lagrange çarpanını, $(0 \le a_i \le C), \quad i = 1,2,...n$

C : DVM hiper parametresini,

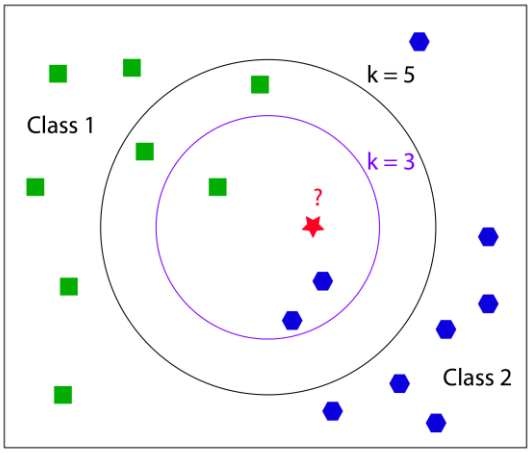
$y_i y_j$: Sınıf verisini,

$K(x_i, x_j)$: Çekirdek fonksiyonunu temsil etmektedir.

3.2.5 k-En Yakın Komşuluk

Örüntü tanımada k-EYK algoriması, sınıflandırmada ve regresyonda kullanılan parametrik olmayan bir yöntemdir. Her iki olguda da giriş, öznitelik uzayındaki k parametresine en yakın eğitim örneklerinden oluşmaktadır (Altman, 1992). K-EYK sınıflandırıcısında, komşuları tarafından oy çokluğuyla sınıflanan bir nesne, k adet en yakın örnekten çevresinde en çok bulanan sınıfa atanır. k parametresinin küçük, tek ve pozitif bir sayıdan oluşması idealdir. Eğer k=1 seçilirse, obje kendisine en yakın olan ilk sınıfa atanır.

Şekil 3.4'de, kırmızı renkle belirtilen test örneğine ait sınıf verisinin tespit edilme yöntemi gösterilmiştir.



Şekil 3.4: k-EYK algoritmasına göre sınıf tayini.

k-EYK algoritması kullanılarak yapılan bu tespit sisteminde k=3 için test örneği kendisine en yakın olan mavi sınıfa atanacaktır. k=3 için bu örnek, kendisine en yakın üç örneğin mavi sınıfa ait olması nedeniyle mavi sınıfa atanacaktır (Mavi: 2, Yeşil: 1). Ancak k=5 değeri için durum daha farklıdır. Test örneğine en yakın beş örnekten çoğunluk bu kez yeşil sınıftan olacağından, örnek yeşil sınıfa atanacaktır (Mavi:2, Yeşil:3). . k-EYK algoritmasın da, k değeri sezgisel olarak belirlenir ve farklı uzaklık ölçütleri kullanılabilir.

Bu çalışmada k-EYK algoritması için öklid uzaklığı kullanılmıştır.

3.3 Başarım Analizi

Sınıflandırıcıların başarımları doğruluk, ROC, F-Ölçütü, kesinlik, MKK ve Sw metrikleri ile değerlendirilmiştir. Bu başarım metrikleri testler sonucunda elde edilen Çizelge 3.2'de görülen karmaşıklık matrisinden hesaplanmaktadır.

Çizelge 3.2: Sınıflandırma Modeli: Karışıklık Matrisi.

	Öngörülen Sınıf		
Gerçek Sınıf		Artı	Eksi
	Artı	Doğru-Artı(DA)	Yanlış-Eksi(YE)
	Eksi	Yanlış-Artı(YA)	Doğru-Eksi(DE)

$Doğruluk$: Elde edilen doğruluk değerini,

DA : Doğru tahmin edilen gerçek pozitifleri,

DE : Doğru tahmin edilen gerçek negatifleri,

YA : Yanlış tahmin edilen pozitifleri (gerçek olmayan pozitifler),

YE : Yanlış tahmin edilen negatifleri (gerçek olmayan negatifler) gösterir.

Doğruluk, sınıflandırma işlemi sonucunda tahmin edilen sonuçların gerçek sonuçlara olan yakınlığını belirtir. Yüzde ile ifade edilir (Taylor, 1998).

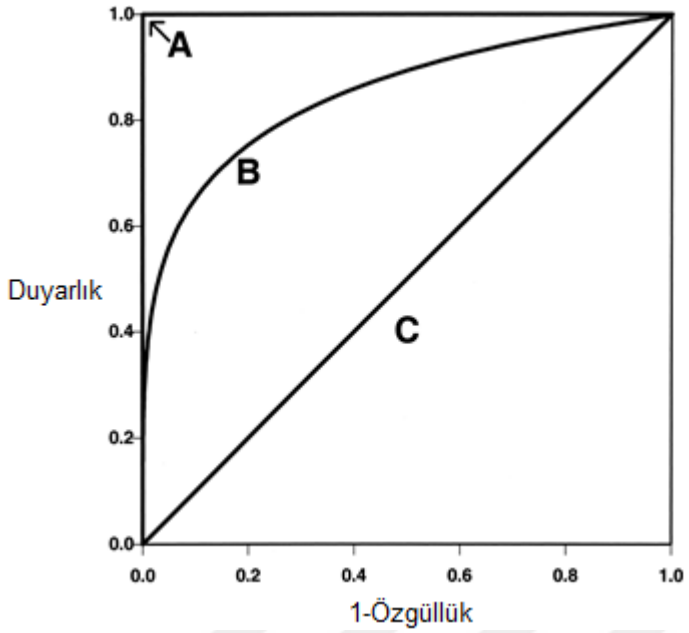
$$Doğruluk = \frac{DA + DE}{DA + YE + YA + DE} \quad \textbf{(3.9)}$$

Model başarımının ölçülmesinde kullanılan en popüler ve basit yöntem, modele ait doğruluk oranıdır. Doğru sınıflandırılmış örnek sayısının ($DA + DE$), toplam örnek sayısına ($DA + YE + YA + DE$) oranıdır.

ROC eğrisi altında kalan alan (AUC: Area Under ROC Curve) ROC eğrisi bir tanı testine ilişkin duyarlık ve özgüllük değerleri arasındaki ilişkiyi grafiksel olarak gösterir (Bradley, 1997). Bu eğrinin altında kalan alanın hesaplanmasıyla AUC elde edilir. AUC, sıralamaya dayalı bir performans kriteridir. AUC, sınıflayıcı modelinin hasta ve sağlıklı kişilerden rastgele seçilmiş iki kişiyi doğru tanımlayabilme olasılığı olarak ifade edilebilir (Vapnik, 1982). 0 ile 1 arasında değerler alabilir; 0.5 değeri rastgele bir tahmin olduğunu, 1'e yakın değerler modelin tahmin gücünün yüksek olduğunu gösterir. ROC eğrisini oluşturan duyarlık ve özgüllük, Eşitlik 3.10 ve Eşitlik 3.11 ile hesaplanır.

$$Duyarlık = \frac{DA}{Tüm\ Pozitifler} \quad \textbf{(3.10)}$$

$$Özgüllük = \frac{YA}{Tüm\ Negatifler} \quad \textbf{(3.11)}$$



Şekil 3.5: ROC karakteristiği (C: Kötü Model, B: İyi Model, A: Mükemmel Model).

Şekilde görüldüğü üzere sınıflandırma sonuçlarına ait ROC grafiklerinde 1'e en yakın değer, en iyi karakteristiği göstermektedir.

Kesinlik, sınıfı pozitif olarak tahminlenmiş Doğru Pozitif örnek sayısının, sınıfı pozitif olarak tahminlenmiş tüm örnek sayısına oranıdır.

$$Kesinlik = \frac{DA}{DA + YA} \tag{3.12}$$

Kesinlik ve duyarlılık ölçütleri tek başına anlamlı bir karşılaştırma sonucu çıkarmamıza yeterli değildir. Her iki ölçütü beraber değerlendirmek daha doğru sonuçlar verir. Bunun için f-ölçütü tanımlanmıştır. F-ölçütü, kesinlik ve duyarlılığın harmonik ortalamasıdır.

$$F - \text{Ölçütü} = \frac{2xDuyarlılıkxKesinlik}{DuyarlılıkxKesinlik} \tag{3.13}$$

Matthews korelasyon katsayısı (MKK), iki sınıflı sınıflama problemlerinde model kalitesini belirten bir ölçüdür. En önemli özelliği sınıflardaki kişi sayıları dengesiz olduğunda diğer kriterlere göre daha doğru sonuç vermesidir. [-1 ile 1] arasında değerler alır. 1 en iyi tahmini, 0 şansa bağlı bir tahmin yapıldığını, -1 ise ters tahmin yapıldığını belirtir.

$$MKK = \frac{(DAxDE) - (YAxYE)}{(DA + YE)(DA + YE)(DE + YA)(DE + YE)} \quad \textbf{(3.14)}$$

Basit çapraz doğrulama yöntemi, 1948 yılında Kurtz tarafından sunulmuştur. 1951 yılında Moiser bu yöntemi geliştirerek çift çapraz doğrulamayı, 1982 yılında da Krus ve arkadaşları tarafından çoklu doğrulama teknikleri geliştirilmiştir (Kurtz, 1948) (D.J. Krus, 1982).

Sınıflandırma modellerinin doğruluğunun test edildiği çapraz doğrulama yönteminde veriler ilk olarak test ve eğitim verisi olarak rastgele olarak bölünmelidir. Eğitim verisi sınıflandırma modelinin kurulumu aşamasında kullanılması sırasında test verisi model kurulumunda ayrı bırakılmalıdır, modelin doğruluğu bu yeni veri seti üzerinden test edilerek modelin uygunluğu test edilir (Bishop, 1995).

Daha basit çapraz doğrulama yöntemlerinde; veri setinin % 5-% 33'lük bir kısmı test verisi olarak ayrılmakta ve ayrılan kısım modelin öğrenme aşamasında eğitim verisi olarak kullanılmamaktadır. Geriye kalan kısım üzerinde ise model kurulmakta ve gerçek değerler ile tahmin değerleri karşılaştırılarak modelin performansı hesaplanmaktadır.

4. GELİŞTİRİLEN ÖZNİTELİK KODLAMA YÖNTEMLERİ

Bu başlık altında düzensiz bölgelerin tespiti için geliştirilen iki farklı kodlama yöntemi yöntemine yer veilmiştir. İlk olarak FKProfil olarak adlandırılan amino asitlerin fizikokimyasal özellikleri, PSM matrisleri, SPINE-X özellikleri kullanılarak oluşturulan yöntemdir.

İkincil olarak bu ilk yöntemde elde edilen özellik vektörleri dalgacık yöntemi ile birlikte kullanılarak yeni özellik vektörleri elde edilmiştir. Bu yöntem ise FKDalgacık olarak adlandırılmıştır.

4.1 FKProfil Öznitelik Kodlama Yöntemi

Protein dizilimlerinden elde edilen başarılı bir model protein yapısının tahmininde önemli bir rol üstlenir. Bu nedenle amino asitler arasındaki komşuluklar bizim için önemli bir bilgi niteliğindedir (Dunker, ve diğerleri, 2001). Bu sebeple *i*'nci aminoasidi çevreleyen aminoasitler birlikte kullanılmalıdır. k pencere uzunluğundaki tüm amino asitler pencere merkezindeki amino asit için bilgi niteliğindedir. Pencerenin tüm protein uzunluğunca kaydırılarak tarama yapılmalıdır. Merkezindeki amino asit pencere içerisindeki ağırlıklı ortalamanın değerini alır (Qian & Sejnowski, 1988). Amino asitleri temsil eden gerçek sayısal değerlerin kullanılması tercih edilmektedir ve pencere içerisindeki tüm amino asitlerin aritmetik ortalaması alınmalıdır (Peng, ve diğerleri, 2005). Bu yol ile örnekteki her bir öznitelik temsil edilmelidir.

Bu çalışmada düzensiz bölgeleri tahmini binary sınıflandırma problemi olarak düşünülmüştür. Makine öğrenmesinde protein içinde bulunan aminoasitlerin sınıf verisi düzensiz bölgeler için "0" düzenli bölge için "1" sayısal değeri alınmıştır. Amino asitler 3 farklı öznitelik türü kullanılarak makine öğrenmesinde kullanılmıştır.

İlk adım olarak kaydırmalı pencereleme metodu ile tüm veri setinden elde edilen sayısal veriler her bir amino asit için uygulanmıştır. Pencerenin merkezindeki aminoasidin alınması için pencere boyutu tek sayı seçilmiştir. Proteindeki ilk aminoasitten başlamak üzere tüm amino asitler için teknik uygulanmalıdır. Bu çalışmada pencere boyutu 101 olarak tanımlanmıştır (Meng, ve diğerleri, 2013). Buradaki temel problem ilk aminoasitten başlandığında, protein dizisi içerisinde başlangıç değerinden önce herhangi bir değer olmaması problem oluşturmaktadır. Bu nedenle başlangıç değerleri *k=0* konumundayken, *w* pencere boyutu olmak üzere (*w-1)/2 - k* adedi kadar önceki değerleri *0* olarak kabul edilmiştir.

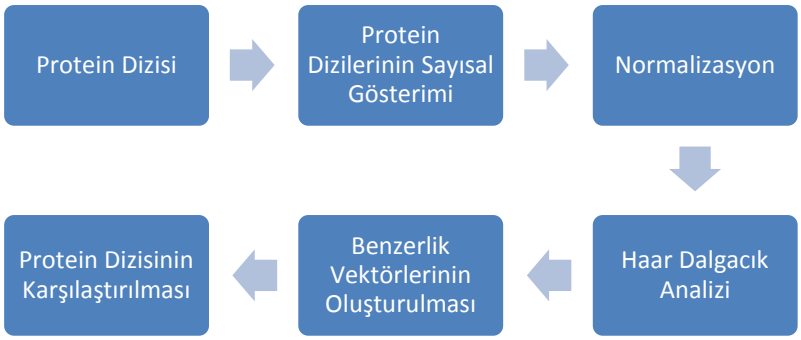
Bir sonraki aşamada her bir örnekten elde edilen öznitelik matrisleri eğitim verisi olarak sınıflandırıcılarda kullanılacaktır.

4.2 FKDalgacık Öznitelik Kodlama Yöntemi

Dalgacık analizi protein sekans analizin de motif araştırması, dizi karşılaştırma problemlerinde uygulanmıştır. Bu bölümde, protein dizi analizinde dalgacık uygulaması anlatılacaktır.

Protein dizisi karşılaştırması biyoinformatik araştırmalarında en önemli alanlarından biridir. Geleneksel BLAST temelli yaklaşım, yerel ikili amino asit eşleşmelerine odaklanır. Ancak, düşük ardışıl özdeş iki protein dizisi, fizikokimyasal özellikleri ve üçüncül yapısında benzerlik gösterebilir ve bu iki protein arasındaki işlevsel bir ilişki olduğunu gösterir.

Farklı çözünürlükte protein dizilerinin karşılaştırılması istenildiği durumlarda dalgacık analizi uygun bir alternatif olmaktadır (de Trad, Fang, & Cosic, 2002). Önerilen yöntemin diyagramı Şekil 4.1 de gösterilmiştir. İlk olarak, protein dizileri kullanılarak sayı dizilerine dönüştürülür. Ardından sayısal dizi sıfır ortalama ve birim standart sapma değerinde normalize edilmelidir. M seviyeli Haar ve bior3.3 dalgacık türleri biorthogonal dalgacık, orthogonal dalgacık ile karşılaştırıldığında daha fazla esneklik sağlamaktadır. Bu nedenle Haar dalgacık modeli tercih edilmiştir. Son olarak, benzerlik vektör katsayıları hesaplanmıştır ve farklı protein dizilerini karşılaştırmak için kullanılmıştır.



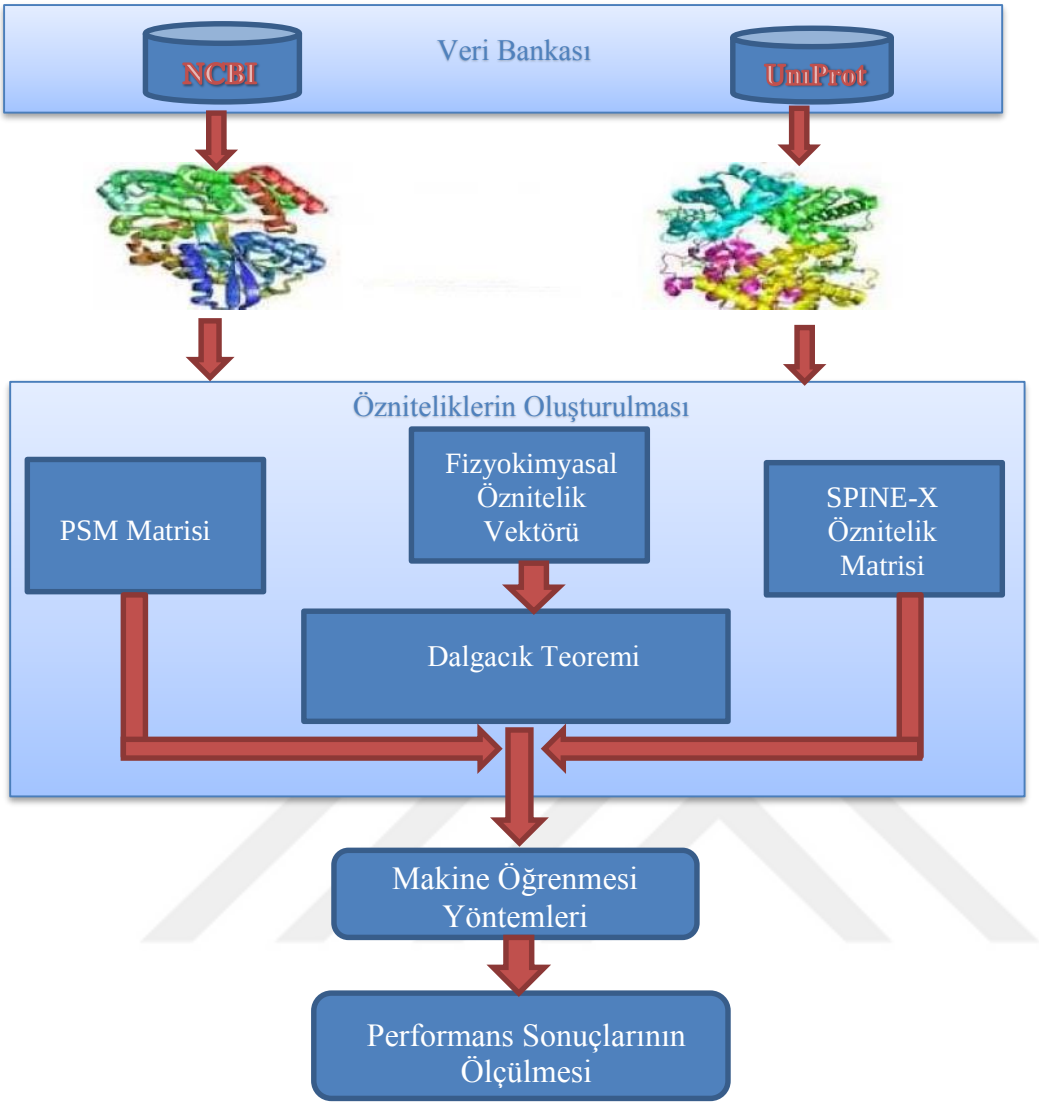
Şekil 4.1: Dalgacık analizi kullanılarak protein dizisinin karşılaştırılması.

Bir sonraki adımda düzenli ve düzensiz proteinler içeren sekanslar PDB' den indirilmiştir. Protein dizilimlerinin anlamlı hale getirilmesi için sayısal değerler ile temsil edilmesi sağlanmıştır. Bu kısımda 7 fizikokimsayal özellik vektörü, AAindex kullanılarak elde edilen PSM Matrisleri ve SPINE-X özellik vektörleri kullanılmıştır. Sonraki adımda 42 özellik vektöründen fizikokimyasal özellikler kullanılarak 7 fizikokimyasal özellik için dalgacık teoremi uygulanarak *M* örnek için *N* dalgacık penceresi kullanılıp *MxN* boyutlarına dönüştürülmüştür. Dönüştürülen öznitelik vektörleri son olarak makine öğrenmesi yöntemlerine uygulanarak düzenli ve düzensiz aminoasitlerin sınıflandırma performansı ölçülmüştür.

Geliştirilen yöntem içerisinde SDR38 ve SDR147 veri setleri kullanılmıştır. Veri setleri içerisinde düzenli ve düzensiz amino asitleri içeren amino asit dizileri bulunmaktadır.

Orijinal protein dizilerine 101 boyutunda bir pencere uygulanmıştır (Meng, ve diğerleri, 2013). Seçilen pencere içerisindeki örnekler dalgacık teoreminde sürekli dalgacık dönüşümü uygulanarak yeni özellik vektörü elde edilmiştir.

Verilerin sürekli dalgacık dönüşümü, Matlab programında Haar özelliği kullanılarak yapılmıştır. Haar dalgacıklar biyolojik dizi analiz uygulamalarında başarılı sonuçlar verdiğinden tercih sebebi olmuştur (Sifuzzaman, Islam, & Ali, 2009) (Meher, Raval, Meher, & Dash, 2012). Deneysel çalışmalarla farklı dalgacıklar arasındaki değişimlere göre skala seviyesi 2 olarak belirlenmiştir. Bu nedenle dalgacık skala seviyesi 1:2:101 olarak belirlenmiştir (Meng, ve diğerleri, 2013). Elde edilen yeni öznitelik vektörleri sınıflandırıcılara uygulanarak test edilmiştir.



Şekil 4.2: FKDalgacık öznitelik kodlama yöntemi.

4.3 Deneysel Sonuçlar ve Analiz

SD38 ve SD147 veri setlerinden FKProfil ve FKDalgacık yöntemine göre öznitelik matrisleri elde edilmiştir. Elde edilen öznitelikler Çizelge 4.1 de görülmektedir.

Çizelge 4.1: FKProfil ve FKDalgacık teoreminden elde edilen öznitelik vektörleri.

Öznitelik Grup ID	Veri Seti	Seçilen Yöntem	Öznitelik Boyutu
1	SDR38	FKProfil	42
2	SDR147	FKProfil	42
3	SDR38	FKDalgacık	749
4	SDR147	FKDalgacık	749

Çizelge 4.1'de düzenli ve düzensiz protein verilerine ait öznitelik vektörleri kullanılarak BayesNet, NaïveBayes ve k-En Yakın Komşuluk sınıflandırıcı modelleriyle 5 katlı çapraz doğrulama yöntemi uygulanarak performans testi yapılmıştır. Elde edilen başarım sonuçları Çizelge 4.2, Çizelge 4.3, Çizelge 4.4'te görülmektedir.

Çizelge 4.2: BayesNet sınıflandırıcısı 5 kat çapraz doğrulama başarım sonuçları.

Öznitelik Grup ID	Doğruluk	AUC	F-Ölçütü	Kesinlik	MKK	Sw
1	85,82%	0,941	0,860	0,866	0,697	0,716
2	81,12%	0,900	0,804	0,817	0,627	0,622
3	81,85%	0,886	0,823	0,841	0,631	0,637
4	77,04%	0,827	0,771	0,777	0,546	0,540

Çizelge 4.3: NaïveBayes sınıflandırıcısı 5 kat çapraz doğrulama başarım sonuçları.

Öznitelik Grup ID	Doğruluk	AUC	F-Ölçütü	Kesinlik	MKK	Sw
1	82.14%	0.907	0.824	0.829	0.615	0,642
2	79,50%	0,864	0,795	0,801	0,595	0,590
3	78,87%	0,811	0,756	0,797	0,519	0,577
4	59,81%	0,711	0,576	0,664	0,271	0,197

Çizelge 4.4: k-EYK sınıflandırıcısı 5 kat çapraz doğrulama başarım sonuçları.

Öznitelik Grup ID	Doğruluk	AUC	F-Ölçütü	Kesinlik	MKK	Sw
1	98,61%	0,991	0,986	0,986	0,969	0,972
2	98,78%	0,990	0,988	0,988	0,975	0,975
3	99,25%	0,992	0,993	0,993	0,983	0,985
4	99,02%	0,990	0,990	0,990	0,980	0,980

Yukarıda elde edilen sonuçlar incelendiğinde FKProfil yöntemiyle elde edilen öznitelikler çapraz doğrulama metodu ile test edildiğinde performans k-En Yakın Komşuluk sınıflandırıcısında % 98,61 doğruluk ile en iyi başarım sonucunun elde edildiği görülmüştür.

Elde edilen yöntemin diğer tahmin araçları ile karşılaştırılması için CASP10 veri seti ile test edilmiştir. Elde edilen başarım sonuçları Çizelge 4.5, Çizelge 4.6, Çizelge 4.7, Çizelge 4.8, Çizelge 4.9' da gösterilmiştir.

Çizelge 4.5: BayesNet sınıflandırıcısı CASP10 veri seti için başarım sonuçları.

Öznitelik Grup ID	Doğruluk	AUC	F-Ölçütü	Kesinlik	MKK	Sw
1	68,22%	**0,583**	0,761	0,886	0,048	0,364
2	75,57%	**0,723**	0,814	0,905	0,175	0,511
3	71,88%	**0,753**	0,789	0,915	0,215	0,437
4	**76,27%**	**0,798**	**0,820**	**0,924**	**0,280**	**0,525**

Çizelge 4.6: NaïveBayes sınıflandırıcısı CASP10 veri seti için başarım sonuçları.

Öznitelik Grup ID	Doğruluk	AUC	F-Ölçütü	Kesinlik	MKK	Sw
1	73,54%	**0,575**	0,796	0,879	0,013	0,470
2	77,85%	**0,756**	0,830	0,913	0,226	0,557
3	76,60%	**0,735**	0,822	0,915	0,235	0,532
4	**79,89%**	**0,795**	**0,925**	**0,925**	**0,308**	**0,597**

Çizelge 4.7: DVM sınıflandırıcısı CASP10 veri seti için başarım sonuçları.

Öznitelik Grup ID	Doğruluk	AUC	F-Ölçütü	Kesinlik	MKK	Sw
1	38,89%	**0,466**	0,507	0,866	-0,035	-0,222
2	74,48%	**0,587**	0,805	0,893	0,101	0,489
3	63,41%	**0,662**	0,726	0,911	0,164	0,268
4	**78,56%**	**0,752**	**0,814**	**0,898**	**0,159**	**0,558**

Çizelge 4.8: k-En Yakın komşuluk sınıflandırıcısı CASP10 veri seti için başarım sonuçları.

Öznitelik Grup ID	Doğruluk	AUC	F-Ölçütü	Kesinlik	MKK	Sw
1	58,84%	**0,556**	0,691	0,888	0,053	0,176
2	64,88%	**0,560**	0,737	0,888	0,059	0,297
3	62,49%	**0,655**	0,719	0,910	0,157	0,249
4	**67,89%**	**0,614**	**0,760**	**0,899**	**0,121**	**0,357**

Çizelge 4.9: SMO sınıflandırıcısı CASP10 veri seti için başarım sonuçları.

Öznitelik Grup ID	Doğruluk	AUC	F-Ölçütü	Kesinlik	MKK	Sw
1	52,80%	**0,562**	0,639	0,892	0,062	0,056
2	**76,04%**	**0,645**	**0,816**	**0,904**	**0,168**	**0,520**
3	54,32%	**0,589**	0,652	0,898	0,089	0,086
4	75.92%	**0,643**	0,816	0,903	0,165	0,518

Sonuçlar incelendiğinde, CASP10 veri seti için en yüksek sonuç NaïveBayes sınıflandırıcısında % 79,89 doğrulukla elde edilmiştir. FKDalgacık yönteminin diğer elde edilen yönteme göre başarılı olduğu söylenebilir.

Çizelge 4.10 ve Çizelge 4.10’ da elde edilen sonuçlar literatürdeki diğer tahmin araçları ile karşılaştırılmıştır (Monastyrskyy, 2014).

Çizelge 4.10: MKK başarım sonuçlarının karşılaştırılması.

Araç İsmi	Protein Sayısı	Kesinlik	Doğruluk	MKK	AUC	Sw
OWL2	90	0,442	0,686	**0,387**	0,821	0,372
Ond-CRF2	92	0,244	0,727	**0,311**	0,814	0,454
FKDalgacık-NaiveBayes	**95**	**0,925**	**0,798**	**0,308**	**0,795**	**0,597**
GSmetadisorder	94	0,228	0,728	**0,300**	0,808	0,456
FKDalgacık-BayesNet	**95**	**0,924**	**0,762**	**0,280**	**0,798**	**0,525**
GSmetaserver	94	0,132	0,699	**0,204**	0,778	0,398
sDispred	94	0,154	0,708	**0,228**	0,778	0,416
GSmetadisorder3d	90	0,261	0,572	**0,173**	0,753	0,144
Slbio	87	0,390	0,687	**0,362**	0,699	0,374
DisMeta	93	0,600	0,692	**0,464**	0,692	0,384
Algorithmic_code	94	0,123	0,599	**0,122**	0,599	0,198
Naïve	94	0,410	0,610	**0,282**	-	0,220

Çizelge 4.11: AUC başarım sonuçlarının karşılaştırılması.

Araç İsmi	Protein Sayısı	Kesinlik	Doğruluk	MKK	AUC	Sw
OWL2	90	0,442	0,686	0,387	**0,821**	0,372
Ond-CRF2	92	0,244	0,727	0,311	**0,814**	0,454
GSmetadisorder	94	0,228	0,728	0,300	**0,808**	0,456
FKDalgacık-BayesNet	**95**	**0,924**	**0,762**	**0,280**	**0,798**	**0,525**
FKDalgacık-NaiveBayes	**95**	**0,925**	**0,798**	**0,308**	**0,795**	**0,597**
GSmetaserver	94	0,132	0,699	0,204	**0,778**	0,398
sDispred	94	0,154	0,708	0,228	**0,778**	0,416
GSmetadisorder3d	90	0,261	0,572	0,173	**0,753**	0,144
Slbio	87	0,390	0,687	0,362	**0,699**	0,374
DisMeta	93	0,600	0,692	0,464	**0,692**	0,384
Algorithmic_code	94	0,123	0,599	0,122	**0,599**	0,198
Naïve	94	0,410	0,610	0,282	-	0,220

5. SONUÇLAR

Elde edilen sonuçlar incelendiğinde CASP10 veri seti için en yüksek doğruluk oranı BayesNet ve NaïveBayes sınıflandırıcıları ile elde edilmiştir. MKK ve AUC ölçütleri dikkate alıdığında BayesNet ve NaïveBayes sınıflandırıcılarının bu problem için uygun bir sınıflandırıcı olduğu görülmüştür. Ayrıca önerilen yöntemlerin diğer yöntemlerle karşılaştırılması için MKK ve AUC değerlerine göre başarı sıralaması yapılmıştır.

Önerilen yöntemlerin başarı sıralamasına göre yapılan değerlendirmede dalgacık teoremi ve yeni profil çıkarım algoritmalarının, düzensiz proteinlerin tahminin de etkili bir yöntem olduğu görülmüştür. FKDalgacık yöntemi ile FKProfil yöntemleri arasında yapılan karşılaştırma neticesinde dalgacık teoremi kullanılarak yapılan öznitelik kodlama yöntemi diğer yönteme göre performansı ciddi oranda arttırmıştır.

Daha sonraki çalışmalarda, geliştirilen algoritmaların ve uygulanan sınıflandırıcı yöntemlerinin bir arada kullanılmasıyla çevrimiçi tahmin aracı oluşturulması amaçlanmaktadır. Çevrimiçi tahmin aracı sayesinde uygulanan yöntemin diğer araştırmacılar tarafından etkin bir şekilde kullanılması sağlanabilir. Web tabanlı bir uygulama yapılması için geliştirilen algoritmanın web tabanlı bir arayüzde tekrardan tasarlanması ve algoritmanın optimize edilerek online kullanılabilir hale getirilmesi gereklidir.

Çalışma da kullanılan veri seti, farklı veri setleri ile karşılaştırılarak performansı arttırmaya yönelik uygun veri seti belirlenebilir. Veri seti içerisindeki uzun ve kısa düzensiz proteinlerin yoğunluk derecelerine göre veri setlerinin performans sonuçları karşılaştırılabilir.

Çalışmada kullanılan 42 özniteliğin tamamı dalgacık teoremine uygulanarak öznitelik uzayı daha çok genişletilebilir. Dalgacık teoremin uygulandıktan sonra en iyi özniteliklerin belirlenmesi mümkündür. En iyi öznitelikleri belirlemek için literatür de kullanılan öznitelik seçme algoritmaları uygulanabilir. Öznitelik seçme algoritmaları sonucunda oluşacak performans sonuçları karşılaştırılarak performasının arttırılması amaçlanmaktadır.

KAYNAKLAR

Alpaydin, E. (2014). Introduction to machine learning. *MIT press*.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*(3), 175-185.

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

Ben-Gal, I. (2007). Bayesian networks. *Encyclopedia of statistics in quality and reliability*.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford university press.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145-1159.

Cai, Y. D., Liu, X. J., Xu, X. B., & Chou, K. C. (2002). Artificial neural network method for predicting protein secondary structure content. *Computers & chemistry, 26*(4), 347-350.

Chang, C. C., & Lin, C. J. (2001). LIBSVM: a library for support vector machines. *LIBSVM: a library for support vector machines*, *2*(3), 27. Software available at http.www.CSIE.NTU.EDU. TW/~ CJLIN/ PAPERS/ LIBSVM. adresinden alındı

Cheng, J., Sweredoski, M. J., & Baldi, P. (2005). Accurate prediction of protein disordered regions by mining protein structure data. *Data mining and knowledge discovery, 11*(3), 213-222.

Chou, P. Y., & Fasman, G. D. (1978). Empirical predictions of protein conformation. *Annual Review Biochemistry, 47*(1), 251-276.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *20*(1), 37-46.

D.J. Krus, E. F. (1982). Computer-Assisted Multicross-Validation in Regression Analysis. *42*, 187-193.

de Trad, C. H., Fang, Q., & Cosic, I. (2002). Protein sequence comparison based on the wavelet transform approach. *Protein engineering, 15*(3), 193-203.

Deng, X., Eickholt, J., & Cheng, J. (2009). PreDisorder: ab initio sequence-based prediction of protein disordered regions. *BMC bioinformatics, 10*(1), 1.

Disfani, F. M., Hsu, W. L., Mizianty, M. J., Oldfield, C. J., Xue, B., Dunker, A. K., & Kurgan, L. (2012). MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics, 28*(12), i75-i83.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification. 2nd. Edition.* John Wiley & Sons Inc,.

Dunker, A. K., Lawso, J., Brown, C. J., Willias, R. M., Romero, P., & Oh, J. S. (2001). Intrinsically Disordered Protein. *Journal of Molecular Graphics and Modeling, 19*, 26-59.

Dunker, A. K., Romero, P., Obradovic, Z., Garner, E. C., & Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. *Genome Informatics, 11*, 161-171.

Eickholt, J., & Cheng, J. (2013). DNdisorder: predicting protein disorder using boosting and deep networks. *BMC bioinformatics, 14*(1), 1.

Faraggi, E., Yang, Y., Zhang, S., & Zhou, Y. (2009). Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. *Structure, 17*(11), 1515-1527.

Fauchere, J. L., Charton, M., Kier, L. B., Verloop, A., & Pliska, V. (1988). Amino acid side chain parameters for correlation studies in biology and pharmacology. *International journal of peptide and protein research, 32*(4), 269-278.

Fischer, E. (1984). Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft, 27*(3), 2985-2993.

Gargour, C., Gabrea, M., Ramachandran, V., & Lina, J. M. (2009). A short introduction to wavelets and their applications. *IEEE circuits and systems magazine, 9*(2), 57-68.

Gwet, K. (2001). Statistical Tables for Inter-Rater Agreement.

Haar, A. (1910). Zur theorie der orthogonalen funktionensysteme. *69*(3), 331-371.

Han, P., Zhang, X., Norton, R. S., & Feng, Z. P. (2006). Predicting disordered regions in proteins based on decision trees of reduced amino acid composition. *Journal of Computational Biology, 13*(10), 1723-1734.

Hirose, S., Shimizu, K., Kanai, S., Kuroda, Y., & Noguchi, T. (2007). POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics, 23*(16), 2046-2053.

İmer, O., & Çavas, X. (2009). The bioinformatics tools for the estimation of disordered regions in proteins. *14th National Biomedical Engineering Meeting, BIYOMUT 2009* (s. 1-4). IEEE.

Ishida, T., & Kinoshita, K. (2007). PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic acids research, 33*(2), W460-W464.

Jiang, Z. F. (2005). Support vector machine for mechanical faults classification. *Journal of Zhejiang University Science*, 433-439.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology, 292*(2), 195-202.

Jones, D. T., & Cozzetto, D. (2015). DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics, 31*(6), 857-863.

Jones, D. T., & Ward, J. J. (2003). Prediction of disordered regions in proteins from position specific score matrices. *Proteins: Structure, Function, and Bioinformatics, 53*(s6), 573-578.

Jones, D. T., & Ward, J. J. (2003). Prediction of disordered regions in proteins from position specific score matrices. *Proteins: Structure, Function, and Bioinformatics, 53*(S6), 573-578.

Kaya, İ. E. (2008). Computational Prediction of Disordered Regions in Proteins. *Department of Electrical and Electronics Engineering*. Adana, Turkey: Çukurova University.

Kohavi, R., & Provost, F. (1998). Glossary of terms. *Machine Learning, 30*(2-3), 271-274.

Kurtz, A. (1948). A Research Test of The Rorschach Test. *Personnel Psychology, 1*(1), 41-51.

Landis, J. R. (1977). The measurement of observer agreement for categorical data. *33*, 159-174.

Li, X., Brown, C. J., Obradovic, Z., Garner, E. C., & Dunker, A. K. (2000). Comparing predictors of disordered protein. *Genome Informatics*(11), 172-184.

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., & Russell, R. B. (2003). Protein Disorder Prediction:Implications for Structural Proteomics. *Structure, 11*(11), 1453-1459.

Mannila, H. (1996). Data mining: machine learning, statistics, and databases. *In ssdbm*, 2.

Meher, J. K., Raval, M. K., Meher, P. K., & Dash, G. N. (2012). Wavelet Transform for Detection of Conserved Motifs in Protein Sequences with Ten Bit Physico-ChemicalProperties. *International Journal of Information and Electronics Engineering, 2*(2), 200-204.

Meiler, J., Müller, M., Zeidler, A., & Schmäschke, F. (2001). Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Article in Journal of Molecular Modeling, 7*(9), 360–369.

Meng, T., Soliman, A. T., Shyu, M. L., Yang, Y., Chen, S. C., & Iyengar, S. S. (2013). Wavelet analysis in current cancer genome research: a survey. *Computational Biology and Bioinformatics, 10*(6), 1442-14359.

Meng, T., Soliman, A. T., Shyu, M. L., Yang, Y., Chen, S. C., Iyengar, S. S., & Iyengar, P. (2013). Wavelet analysis in current cancer genome research: a survey. *Computational Biology and Bioinformatics, 10*(6), 1442-14359.

Monastyrskyy, B. K. (2014). Assessment of protein disorder region predictions in CASP10. *Proteins: Structure, Function, and Bioinformatics, 82*(S2), 127-137.

Mount, D. W. (2001). *Bioinformatics: sequence and genome analysis* (Cilt 2). NewYork: Cold spring harbor laboratory press.

Možina, M., Demšar, J., Kattan, M., & Zupan, B. (2004). *Nomograms for visualization of naive Bayesian classifier.* Springer Berlin Heidelberg.

N. Friedman, M. G. (1997). Challenge: what is the impact of Bayesian Networks on learning? *15th international joint conference on artifical intelligence. 1*, s. 10-15. Morgan Kaufmann Publishers Inc.

Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., Dunker, A. K., & Obradovic, Z. (2005). Optimizing long intrinsic disorder predictors with protein evolutionary information. *Journal of bioinformatics and computational biology*, *3*(1), 35-60.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

Polatkan, A. C. (2007). Protein Homoloji Tespitinde Bir Üst Sınıflandırma Yaklaşımı. *Bilgisayar Mühendisliği Anabilim Dalı*. Ankara: Başkent Üniversitesi.

Powers, D. (2012). The Problem of Kappa. *13thConference of the European Chapter of the Association for Computational Linguistics*, (s. 345–355).

Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E. H., Man, O., Beckmann, J. S., & Sussman, J. L. (2005). FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, *21*(16), 3435-3438.

Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of molecular biology*, *202*(4), 865-884.

Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of molecular biology*, *202*(4), 865-884.

***RCSB Protein Data Bank*.** (2016, 3 5). 2016 tarihinde http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=molType-protein&seqid=100 adresinden alındı

Rifkin, R. M. (2002). Everything old is new again: a fresh look at historical approaches in machine learning. MaSSachuSettS InStitute of Technology.

Riis, S. K., & Krogh, A. (1996). Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *Journal of Computational Biology*, *3*(1), 163-183.

Romero, P., Obradović, Z., Kissinger, C., Villafranca, J. E., & Dunker, A. K. (1997). Identifying disordered regions in proteins from amino acid sequence. *Neural Networks* (s. 90-95). International Conference on IEEE.

Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., & Dunker, A. K. (2001). Sequence complexity of disordered protein. *Proteins: Structure, Function, and Bioinformatics*, *42*(1), 38-48.

Rost, B., & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of molecular biology*, *232*(2), 584-599.

Sifuzzaman, M., Islam, M. R., & Ali, M. Z. (2009). Application of Wavelet Transform and Its Advantages Compared to Fourier Transform. *Journal Physical Sciences*(13), 121-134.

Sifuzzaman, M., Islam, M. R., & Ali, M. Z. (2009). Application of Wavelet Transform and Its Advantages Compared to Fourier Transform. *Journal Physical Sciences, 13*, 121-134.

Simon, P. (2013). *Too Big to Ignore: The Business Case for Big Data.* John Wiley & Sons.

Sirota, F. L., Ooi, H. S., Gattermayer, T., Schneider, G., Eisenhaber, F., & Maurer-Stroh, S. (2010). Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. *BMC genomics, 11*(1), 1.

Stormo, G. D., Schneider, T. D., Gold, L., & Ehrenfeucht, A. (1982). Use of the 'Perceptron'algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research, 10*(9), 2997-3011.

Swets, J. A. (1996). Signal detection theory and ROC analysis in psychology and diagnostics. *Lawrence Erlbaum Associates*.

Taylor, R. (1998). An Introduction to Error Analysis: The Study of Uncertainties in PhYSical Measurements. 128-129.

Vapnik, V. N. (1982). *Estimation of Dependences Based on Empirical Data.* New York: Springer-Verlag.

Vullo, A., Bortolami, O., Pollastri, G., & Tosatto, S. C. (2006). Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic acids research, 34*(2), W164-W168.

Wang, L., & Sauer, U. H. (2008). OnD-CRF: predicting order and disorder in proteins conditional random fields. *Bioinformatics, 24*(11), 1401-1402.

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

Wernick, M. N., Yang, Y., Brankov, J. G., Yourganov, G., & Strother, S. C. (2010). Machine learning in medical imaging. (IEEE, Dü.) *Signal Processing Magazine, 27*(4), 25-38.

Wlodawer, A., & Erickson, J. W. (1993). Structure-based inhibitors of HIV-1 protease. *Annual review of biochemistry, 62*(1), 543-585.

Wu, C. H., & McLarty, J. W. (2012). *Neural networks and genome informatics (Vol. 1).* USA: Elseiver.

Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K., & Uversky, V. N. (2010). PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1804*(4), 996-1010.

Yang, Z. R., Thomson, R., McNeil, P., & Esnouf, R. M. (2005). RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics, 21*(16), 3369-3376.

Zanni, L., Serafini, T., & Zanghirati, G. (2006). Parallel software for training large scale support vector machines on multiprocessor systems. *The Journal of Machine Learning Research, 7*, 1467-1492.

Zhang, H. (2005). Exploring conditions for the optimality of naive Bayes. *International Journal of Pattern Recognition and Artificial Intelligence, 19*(02), 183-198.

Zhang, T., Faraggi, E., & Zhou, Y. (2010). Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction. *Proteins: Structure, Function, and Bioinformatics, 78*(16), 3353-3362.

Zhang, T., Faraggi, E., Xue, B., Dunker, A. K., Uversky, V. N., & Zhou, Y. (2012). SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. *Journal of Biomolecular Structure and Dynamics, 29*(4), 799-813.

ÖZGEÇMİŞ

Ad Soyad:	Sebahattin BABUR
Doğum Yeri ve Tarihi:	Bursa / 04.08.1988
Adres:	Gedik Üniversitesi, Meslek Yüksek Okulu, Tıbbi Görüntüleme Teknikleri Programı
E-Posta:	sebahattin.babur@gedik.edu.tr
Lisans:	Marmara Üniversitesi, Elektronik Bilgisayar Eğitimi Bölümü, Elektronik ve Haberleşme Öğretmenliği Programı

Mesleki Deneyim ve Ödüller:

2015 - Halen	Öğretim Görevlisi, Gedik Üniversitesi, Meslek Yüksek Okulu, Tıbbi Görüntüleme Teknikleri Programı

Yayın ve Patent Listesi:

- Turhal, U., Gök M., Onur S., **Babur S.,** "Performance Analysis of Feature Ranking Algorithms on Microarray Datasets". *The 5th International Symposium on Sustainable Development (ISSD 2014),* Sarajevo - Bosnia & Herzegovina, 15-18 May 2014.

- Akbaş A., Turhal U., **Babur S.** and Avcı C.,"Performance Improvement with Combining Multiple Approaches to Diagnosis of Thyroid Cancer". *The 7th International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2013),* Beijing - China, 26-28 Sep 2013.

- Turhal U., **Babur S.,** Avcı C., Akbas A., "Performance Improvement for Diagnosis of Colon Cancer by Using Ensemble Classification Methods". *The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE 2013),* Konya - Turkey, 9-11 May 2013.

- Böcekci.S.,**Babur,S.,**Böcekci,G.,Baba,F.,"Akış ve Sıvı Seviye Kontrol Siste mi için PAC Tabanli SCADA Uygulaması ",
Turkish National Meeting on Automatic Control (TOK 2013)

- **Babur S.,** Turhal U., Akbaş A., "Dvm Tabanlı Kalın Bağırsak Kanseri Tanısı İçin Performans Geliştirme". *Elektrik-Elektronik ve Bilgisayar Mühendisliği Sempozyumu (ELECO 2012)*, Bursa - Türkiye, 29 Kasım-1 Aralık 2012.

TEZDEN TÜRETİLEN YAYINLAR/SUNUMLAR

- **Babur, S.,** Gök, M., 2016: Prediction of Disordered Protein Regions with Voting Ensemble Classification Method (Düzensiz Protein Bölgelerinin Birleştirilmiş Oylama Sınıflandırma Yöntemi ile Tahmini). *International Conference on Research in Education & Science(ICRES),* Mayıs 19-22, 2016 Bodrum, Turkey