Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Biyolojik Protein Fonksiyon Tahmin İşleminde Orange Veri Analizi Aracının Kullanımıyla Makine Öğrenmesi Algoritmalarının Performanslarının Değerlendirilmesi

Ceren AKMAN YAMAN [a,*], Volkan ALTUNTAŞ [b]

[a] *Bilgisayar Mühendisliği Bölümü, Mühendislik Mimarlık Fakültesi, Kırşehir Ahi Evran Üniversitesi, Kırşehir, TÜRKİYE*

[b] *Bilgisayar Mühendisliği Bölümü, Mühendislik ve Doğa Bilimleri Fakültesi, Bursa Teknik Üniversitesi, Bursa, TÜRKİYE*

* *Sorumlu yazarın e-posta adresi: ceren.akman@ahievran.edu.tr*

DOI: 10.29130/dubited.1162917

Öz

İnsan vücudu ilk günden bugüne kadar olan bütün süreçlerde işleyiş açısından merak uyandıran bir mekanizma olmuştur. İçerisinde barındırdığı hücrelerle bu hücrelerin kendi içlerinde barındırdıkları moleküllerle ve işleyişlerle yaşamsal döngü devam etmiştir ve devam etmektedir. Bu yaşamsal döngünün devam etmesi için moleküllerin işleyiş şekillerinin anlaşılmasının yaşamsal faaliyetlerin çözümlenmesinde önemli etkisi olduğu kanısına varılmıştır. Bu çalışma kapsamında yapılan çalışmalar incelendiğinde insan vücudu için karmaşık bir yapıya sahip olan moleküllerin işleyişinin büyük bir öneme sahip olduğu kararına varılmıştır. Böylelikle bu çalışma da büyük bir öneme sahip olan karmaşık yapılı protein molekülü ele alınarak biyoloji tarafından bakıldığında biyolojik süreç, moleküler işlev ve hücresel bileşen açısından fonksiyon tahmin işleminin gerçekleştirilebilmesi ve bunun için k- en yakın komşuluk, sinir ağı ve rastgele orman yöntemlerinin veri görselleştirme ve veri analiz aşamasında kullanılabilen Orange editörü vasıtasıyla modellerin geliştirilmesi sağlanmış olup performans değerlendirilmesi yapılmıştır. Yapılan değerlendirmeler sonucunda k-en yakın komşuluk modelinin kullanılan veri setleri üzerinde en az %88 üzerinde başarı sağladığı tespit edilmiştir.

Anahtar Kelimeler: *Biyoinformatik, Protein fonksiyonu, Sinir ağı, Rastgele orman, K-En yakın komşuluk*

Evaluation of Performance of Machine Learning Algorithms Using Orange Data Analysis Tool in Protein Function Estimation Process

Abstract

The human body has been an intriguing mechanism in terms of functioning in all processes from the first day to the present. The vital cycle has continued and continues with the cells it contains and the molecules and processes that these cells contain. It has been concluded that understanding the functioning of molecules has an important effect on the analysis of vital activities in order to continue this vital cycle. When the studies carried out within the scope of this study were examined, it was concluded that the functioning of molecules, which have a complex structure for the human body, is of great importance. Thus, by considering the complex protein molecule, which is of great importance in this study, it is possible to perform the function estimation process in terms of biological process, molecular function and cellular component, and for this, data visualization and data visualization of k-nearest neighbor, neural network and random forest methods. The development of the models was provided by the Orange editor, which can be used in the analysis phase, and the performance evaluation was made. As a result of

Geliş: 16/08/2022, Düzeltme: 12/10/2023, Kabul: 13/10/2023

the evaluations, it was determined that the k-nearest neighbor model achieved at least 88% success on the data sets used.

Keywords: *Bioinformatics, Protein function , Neural network, Random forest, K-Nearest neighborhood*

I. GİRİŞ

Proteinler, aminoasitlerin zincir şeklinde bağlanması sonucu oluşan ve içinde azot barındıran temel yapısal polimerlerdir. Büyük bir öneme sahip olan bu polimerler için aynı zamanda bilinen en karmaşık yapılı moleküller olduğu söylenmektedir. Bu moleküller tüm biyolojik dokuların ve hücrelerin hayati yani ana yapısal bileşeni olarak ve yaşam aktivitelerinin birincil taşıyıcısı şeklinde tanımlanmaktadır [1]–[3] . Vücuttaki kasların, dokuların ana yapısal bileşeni olduğu bilinen proteinler hormonları, enzimleri üretmek, dokuyu inşa etmek, onarmak, bağışıklık sistemi güçlendirmek, azotu vücutta kullanılabilir hale getirmek, kas gelişimini sağlamak gibi görevleri bulunmaktadır. Ancak bu moleküllerin vücut tarafından bu veyahut başka şekillerde kullanılabilmesi için en temel halleri olan aminoasitlere kimyasal dönüşümlerinin gerçekleşmesi gerekmektedir. Canlı bünyesi için gerekli olan aminoasitlerin doğada güncel olarak 20 çeşit olduğu ve bunların kendi içlerinde özel dizilimleri olduğu söylenmektedir. Bu özel dizilimler sayesinde proteinlerin işlevleri belirlenmektedir. Bu işlevlerin veyahut dizilimlerin belirlenmesi hastalık teşhisi ve tedavisi, ilaç keşfi gibi hem biyoloji alanı için hem de tıp ve eczacılık alanı için büyük bir önem arz etmektedir [1], [4], [5].

Pitre ve arkadaşları tarafından gerçekleştirilen çalışmada, hücre içindeki çalışmaların büyük bir kısmının proteinler tarafından yapıldığı söylenmektedir. Böylelikle hücrelerin davranış ve çeşitli reaksiyonlarına karşı cevaplarından protein-protein etkileşimlerinin neden olduğu söylenmiştir [6]. Canlı vücudundaki biyolojik reaksiyonların birçoğu proteinlerin birbirleri ile etkileşimleri yani bağlanıp veya ayrışmaları sonucunda oluşmaktadır. Bu etkileşimlerin oluşmasını sağlayan nedenleri anlamak moleküler düzeyde biyolojik olarak önemli olayların kontrolünü sağlamak açısından büyük öneme sahip olduğu bilinmektedir. Bu nedenle proteinler arasındaki etkileşimi tahmin edebilmek ya da anlayabilmek açısından literatürde hesaplamalı veya deneysel yöntemler olmak üzere iki farklı yöntem bulunmaktadır. Bu bilgi ışığında yapılan incelemelerde Fukuhura ve arkadaşlarının yapmış oldukları çalışma da proteinlerin etkileşimlerini üç boyutlu yapısını çözebilmek adına birçok deneysel yöntemin olduğu belirtilmiştir [7]. Bununla birlikte Shen ve arkadaşları tarafından gerçekleştirilen çalışmadaysa deneysel yöntemlerle ulaşılan verilerin toplam protein etkileşim ağının sadece bir kısmını oluşturduğu söylenmiştir [8].

Protein – protein etkileşimlerinin ve fonksiyonlarının işleyişlerinin çözümlenebilmesi için laboratuvar ortamında (in-vitro) ve canlı üzerinde (in-vivo) yapılan deneysel işlemler sonucunda etkileşim/fonksiyon tahmin tespit işlemleri gerçekleştirilebilmektedir [4]. Bu bilgi doğrultusunda yapılan araştırmalar sonucunda Cai ve arkadaşlarının gerçekleştirdiği çalışmada protein fonksiyonlarının deneysel yöntemlerle tahmin işleminin gerçekleştirilmesinin uzun zaman alacağı ve yüksek maliyetli olabileceği söylenmiştir. Ayrıca bununla birlikte fonksiyon/etkileşim tahmin işleminde farklı deneysel yöntemlerin kullanılması sonucunda istenmeyen yanlış olarak tanımlanan yanlış-pozitif (false - positive) ya da yanlış-negatif (false - nagative) gibi tespitlerin olabileceği söylenmiştir [9]. Dolayısıyla hem zaman hem de maliyet açısından araştırmacılar için dezavantaj olan deneysel yöntemler yerine son yıllarda etkileşim/fonksiyon tahmin işlemlerinde bilgisayar yardımlı analizleri, modellemeleri kapsayan hesaplamalı yöntemler kullanılmaya başlanmıştır. Bununla birlikte hesaplamalı yöntemlerin bir alt parçası olan ve günümüzde çoğu biyolojik araştırmalardan elde edilen ham verilerin işlenmesini, düzenlenmesini, anlamlandırılmasını sağlayan bunu matematik ve istatistik ile birleştiren bir teknoloji olan biyoinformatikte biyoloji alanında sıkça kullanılmaya başlanmıştır [10], [11].

Biyoloji alanındaki verilerin büyük olması ve her geçen gün bu veri sayılarının kümülatif bir şekilde artmaya devam etmesi makine öğrenmesi teknolojisinin de kullanılmasına neden olmuştur. Bu nedenle

hesaplamalı yöntemlerin kullanımında tercih edilen ilk başlıklardan birinin makine öğrenmesi olduğu söylenmiştir [11]. Makine öğrenmesinin belirli teknikleri kullanılarak biyolojik verilerden bilgi kümeleri çıkartılabilmesi ve daha sonrasında bu veri kümelerini kullanılarak bir model oluşturması biyoinformatik açısından büyük öneme sahiptir. Çünkü biyoinformatikte elde edilen modellerle veriler öğrenilerek bunlar arasındaki ilişkiler ileriye yönelik tahmin edilebilmektedir [12]. Bu bilgiler ışığında protein fonksiyonlarının etkileşimlerinin tahmin edilmesi hem ilaç keşfinde hem de tedavi alanı gibi birçok alanda işlemlerin kolaylaşmasını sağlayacağı söylenebilmektedir.

Bu bilgiler doğrultusunda yapılan incelemeler makine öğrenmesi tekniklerinin protein işlevlerinin tahmininde son zamanlarda sıkça kullanıldığını göstermektedir. Bu kapsamda yapılan incelemeler de Hakala ve arkadaşlarının yaptıkları çalışma da protein fonksiyonları için ileri beslemeli sinir ağı, evrişimsel sinir ağı ve rastgele orman modelleri kullanılarak tahmin işlemi gerçekleştirilmiştir. Bu çalışmanın sonucunda geliştirilen modelin f skor, hassasiyet ve kesinlik ölçümleri yapılmış ve sırasıyla 49%, 45% ve 55% elde edildiği görülmüştür [5]. Kulmanov ve arkadaşları tarafından evrişimli sinir ağı (CNN) modelinin kullanıldığı DeepGOPlus aracı geliştirilmiştir. Böylelikle CNN modelin kullanıldığı bu araçta proteinlerin aminoasit dizisi kullanılarak fmax 47% oranında biyolojik süreçlerin tahmin edilmesi sağlanmıştır. Sonuç olarak DeepGoPlus'a saniyede yaklaşık olarak 40 protein dizisine açıklama ekleyebileceği ve böylelikle çok çeşitli proteinler için hızlı ve doğru tahmin sağlayacağı söylenmektedir [13]. Teknolojinin gelişmesiyle birlikte ham veri sayısı ve içerik karmaşıklığı gün geçtikçe kümülatif bir şekilde artmaktadır. Son zamanlarda adından sıkça söz ettiren derin öğrenme modellerinin karmaşık veriler üzerinde özellikleri belirleme de son derece mükemmel olduğu söylenmektedir [14]. Gelman ve arkadaşlarının gerçekleştirdiği çalışama da derin öğrenme modeli, protein dizisindeki mutasyonel bölgelerin paralel fonksiyonel analizi için deneysel bir çalışma olan derin mutasyonel taramadan elde edilen verilerin kullanılmasıyla protein dizisi ve fonksiyon arasındaki ilişkiyi tahmin etmek için tercih edilmiştir [15].

Büyük ölçekteki proteinlerde fonksiyon tahmini yapmak yaşamın moleküler mekanizmasını anlamak için gerekli olduğu söylenebilmektedir. Bununla birlikte UniProtKB'deki 179 milyondan fazla proteinin yalnızca çok küçük bir yüzdesinin deneysel kanıtlarla desteklenen Gen Ontolojisi (GO) açıklamalarına sahip olduğu bilinmektedir. 1998'de başlatılan bu Gen Ontolojisi, Biyoinformatik alanında yaygın olarak kullanılmaktadır ve GO'nun asıl amacı, terminoloji açıklaması veya gen kelimelerinin ve gen ürün özelliklerinin yorumlanması için temsili bir platform sağlamaktadır. Biyoinformatik araştırmacılarının gen ve gen ürünleri verilerini özetlemelerini, işlemelerini, yorumlamalarını ve paylaşmalarını sağlar. Gen Ontolojisi, Directed Acyclic Graph (DAG) tipi bir ontolojidir. Böylelikle GO, işlevleri ve hücre konumlarını içeren 45.000'den fazla biyolojik kavram içeren ve biyolojinin üç yönünü kapsayan biyolojik süreç (BS), moleküler işlev (Mİ) ve hücresel bileşen (HC) üç kategoriye ayrılır. Bir proteinin genellikle birden fazla GO notu bulunduğundan dolayı protein fonksiyon tahmini çok büyük ölçekli çok etiketli bir sınıflandırma problemidir ve proteinlere GO terimlerini doğru bir şekilde atamak zorlu bir iş olduğu bilinmektedir [2]. Araştırmaları yapılan diğer literatür incelemeleri sonucunda elde edilen referans çalışmalar Tablo 1'de gösterilmektedir.

Tablo 1. *İncelenen çalışmalarla ilgili özet bilgilerin devamını içeren tablosal gösterimi.*

No	Çalışma Adı	Çalışma Kapsamı	Kullanılan Yöntem	Elde Edilen Başarı Oranları
1	Derin Öğrenme Kullanarak Protein İkincil Yapı Tahmini [28]	Proteinlerin canlı organizmalarının en önemli parçalarından biri olduğu aynı zamanda bu yapı taşlarının işlevlerinin ve yapılarının anlaşılmasının büyük önem arz ettiği söylenmiştir. Bu amaca yönelik gerçekleştirilen çalışma da CB513 veri seti üzerinde CNN, RNN, LSTM ve GRU gibi derin öğrenme modelleri kullanılarak karşılaştırma yapılmıştır.	Evrişimli Sinir Ağı(CNN) Yinelemeli Sinir Ağı (RNN) Uzun Kısa Süreli Bellek (LSTM) Kapılı Tekrarlayan Hücre(GRU)	CNN- %82.54 RNN- %81.06 LSTM- %81.10 GRU- %81.48
2	Protein İkincil Yapı Tahmini İçin Makine Öğrenmesi Yöntemlerinin Karşılaştırılması [29]	Protein veritabanlarındaki verilerin hızlı artışının protein yapı tahminini önemli kıldığı bu nedenle de protein ikincil yapı üzerinde tahmin işlemlerinin makine öğrenmesi teknikleriyle gerçekleştirilebileceği söylenmiştir. Bu çalışma kapsamında EVASET veri seti üzerinde makine öğrenmesi modelleri karşılaştırılmıştır. Bunun sonucunda en iyi sonuç SVM ile elde edilmiştir.	Ekstra Öğrenme Makineleri (ELM) K- En Yakın Komşuluk (KNN) Rastgele Orman (RO) Yapay Sinir Ağı (YSA) Destek Vektör Makineleri (DVM)	ELM- %82.74 KNN- %82.10 RO- %83.18 YSA- %83.08 DVM- %83.45
3	SDN2GO: An Integrated Deep Learning Model for Protein Function Prediction [2]	Protein fonksiyonlarını tahmin etmek için SDN2GO isimli entegre bir derin öğrenme tabanlı bir sınıflandırma modeli önerilmiştir. Bu modelde CNN ve ağırlık sınıflandırıcısı kullanılmıştır. Bu çalışma kapsamında her alt ontoloji için test işlemlerinin gerçekleştirilebilmesi adına eğitim setinde 5 kat çapraz doğulama kullanılmıştır. Bunun sonucunda referans alınan BS, Mİ, HC kategorileri için etkili sonuçlar elde edildiği söylenmiştir.	CNN - Ağırlık Sınıflandırma	(İnsan) BP kategorisi için CNN- %92.1 Mİ kategorisi için CNN- %95.7 HC kategorisi için CNN- %95.2 (Maya) BS kategorisi için CNN- %83.9 Mİ kategorisi için CNN- %90.3 HC kategorisi için CNN- %87.8

Tablo 2 (devamı). *İncelenen çalışmalarla ilgili özet bilgilerin devamını içeren tablosal gösterimi.*

4	SVM-Prot 2016: A Web-Server for Machine Learning Prediction of Protein Functional Families from Sequence Irrespective of Similarity [30]	Protein işlevleri tıbbi ve biyolojik bakımdan çok büyük bir öeneme sahip olduğu bu nedenle fonksiyonel tahmin yöntemlerinin geliştirilmesine ihtiyaç duyulduğu söylenmiştir. Bu çalışma kapsamında protein işlevsel ailelerinin üzerinde makine öğrenmesi yöntemleri kullanılarak tahmin işlemleri yapılmıştır.	DVM KNN Olasılıksal Sinir Ağı (PNN) Bu makale kapsamında yapılan değerlendirme kıstaslarında Li ve arkadaşlarının kullanıdığı yöntemlerin kıyaslanması için veri setlerinin duyarlılık sonuçları referans alınmıştır.	Başarı performansları DVM- %50~99.99 KNN- %51.06~99.99 PNN- %60.49~99.99 aralığındadır.
5	Machine learning techniques for protein function prediction [31]	Bu çalışma kapsamında en yaygın taksonomiler üzerinde makine öğrenmesi algoritmalarının proteinlerin işlevlerini tahmin etme işlemlerinin karşılaştırılması yapılmıştır. Bu çalışma kapsamında 3 farklı en yaygın taksonomi referans alınmıştır ancak bu makale kapsamında GO Ontolojisinden elde edilen sonuçlar değerlendirildiği için elde edilen sonuçlar belirtilmiştir.	—	Derin Öğrenme modellerinin GO Ontolojisi üzerinde daha faydalı sonuçlar verdiği belirtilmişir.
6	DEEPred: Automated Protein Function Prediction with Multi-task Feed-forward Deep Neural Networks [32]	DEEPred GO tabanlı protein fonksiyon tahmininde bir çözüm olarak çok görevli ileri beslemeli derin sinir ağlarının hiyerarşik yığını olarak önerilmektedir.	Modellerin performansı kaynak eğitim verilerinin %80 -%20 ayrımıyla çapraz doğrulama kullanılarak ölçüm gerçekleştirilmişitr. Bunun sonucunda GO Ontolojisi kategorilerinden olan Mİ, BS ve HC kategorileri için performans değerlendirilmesi yapılmıştır.	Performans ölçümleri DNN - Mİ %67 DNN - BS %51 DNN - HC %58 olarak tespit edilmiştir.

II. MATERYAL VE METOT

Bu bölümde, protein veri kümelerini, kategorilerini oluşturmak için kullanılan dizi analizleri için Cai ve arkadaşları tarafından [2] gerçekleştirilen çalışmada sunulan veri seti kullanılarak k- en yakın komşuluk, rastgele orman ve yapay sinir ağı sınıflandırıcılarının ayrıntıları ele alınmıştır. Bu veri kümesinde eğitim amaçlı olan veri seti kısmı;

- Deney için UniProt veri tabanından araştırma için gerekli olan proteinlerin sekans bilgileri FASTA formatındaki dosyalar şeklinde indirilerek,

- Daha sonraki aşama da elde edilen bu dosyalar içerisindeki gürültü, fazlalık ya da gereksiz bilgi olarak nitelendirebileceğimiz verilerin kaldırılması için CD-hit aracının kullanımı gerçekleştirilerek,

- Bu araç sayesinde dizi benzerliği 60% üzerinde olan proteinler tek bir grup altında toplanmıştır ve sonrasında benzer olan gruplar sadece 1 protein başlığı altında tutularak, elde edilmiştir.

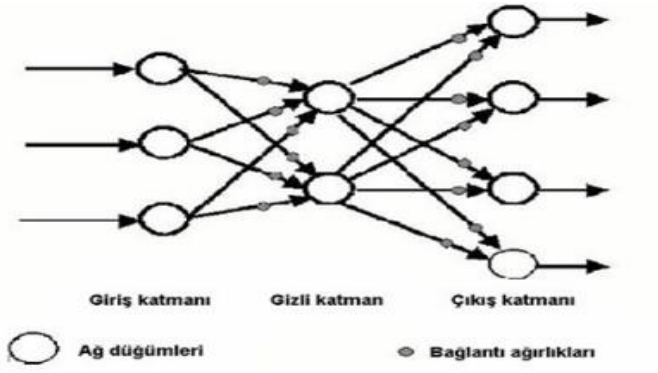
Elde edilen bilgiler doğrultusunda bu çalışma kapsamında sadece insanlar için olan veri setleri değerlendirmeye alınmıştır. Bu nedenle ilgili veri seti içerisinde yukarıda bahsi geçen düzenlemelerden sonra insanlar için 13.704 veri içerik elde edildiği söylenmiştir. Elde edilen bu içerikler için GO açıklama verileri tespit edilmiştir. Böylelikle BS, HC ve Mİ olmak üzere üç kategori başlığı altında toplamda 13.882 içerik olduğu söylenmiştir. Bu işlemlerin ardından veri seti performansının arttırılması için STRING veri tabanı v10'dan türetilen protein-protein etkileşimi (PPI) ağ verileri eklenmiştir. Akabinde tüm UnitProtKB proteinlerini ve InterPro girişlerini ve bunların eşleştiği bireysel imzaları içeren halka açık olan ve proteinlerin ilgili alan verilerini içinde barındıran interpro veri tabanından verilerin elde edilmiştir. Ardından ihtiyaç duyulan proteinler için UnitProt kimliği ile arama yapılarak, gereksiz olabilecek verilerin kaldırılması ya da çelişkili kanıtlarla desteklenen aynı verilerden sadece birinin tutulması sağlanarak veri setinin düzenliliği sağlanmıştır.

Böylelikle bu çalışma kapsamında da düzenliliği sağlanan veri seti üzerinde işlemler gerçekleştirilerek bir sonraki aşama da sinir ağları (SA), rastgele orman (RO) ve k-en yakın komşuluk (KNN) algoritmaları değerlendirilmiştir.

A. SİNİR AĞLARI (SA)

Sinir ağı, biyolojik nöronların bir ağı veya bir devresi olan yapay sinir ağları olarak anılan, makine öğrenmesinin bir alt kümesi olan ve derin öğrenme algoritmalarının oluşumunda yer alan bir modeldir. Yapay sinir ağları öğrenme, düşünme, hatırlama gibi tüm insan davranışlarını taklit etmektedir. Böylelikle yapay sinir ağlarının temelde insan beyninden esinlenerek geliştirildiği söylenebilmektedir [16]. Yapay sinir ağları biyoloji alanında yer alan sinir ağlarından esinlenerek geliştirilmiştir. Böylelikle biyolojik sinaptik ağırlıkları olan nöronların sinyal gönderme durumunu taklit ettiği söylenebilmektedir. Yapay sinir ağları bir girdi katmanı, bir veya birden fazla gizli katman ve bir çıktı katmanı olan düğümlerden oluşmaktadır (Şekil 1). Bu modelin çalışma mantığında her nöron ya da yapay sinir diğer nörona bağlanarak bir ağırlık ve eşiğe sahip olmaktadır. Daha sonrasında herhangi bir nöronun çıktısı eşik değerinin üzerindeyse o nöron etkinleştirilir ve ağın bir sonraki aşamasına gönderilir. Eğer eşik değerinin altındaysa o zaman ağın bir sonraki katmanına iletim gerçekleştirilmez [16], [17].

Yapay sinir ağının karmaşık problemleri çözebilme yeteneğinden dolayı ve tahmine dayalı modelleme de başarılı olacağından bu çalışma kapsamında referans alınan veri setinin özellikleri de göz önünde bulundurulduğunda sinir ağının kullanımı tercih edilmiştir.

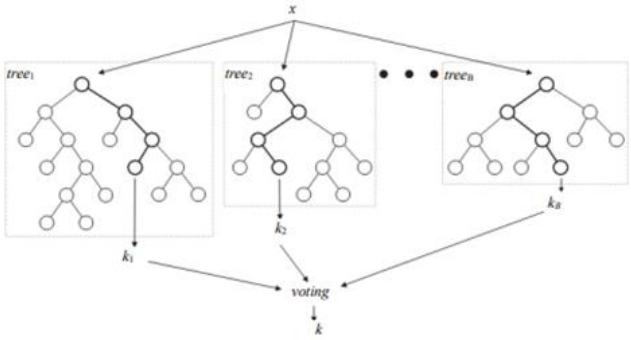


***Şekil 1**. Sinir ağı yapısı (Neural network structure) [16].*

Bu çalışma kapsamında geliştirilen modelde kullanılan sırasıyla 100 ve 50 olmak üzere nöron sayısı içeren 2 adet gizli katman kullanılmıştır. Gizli katmandan çıkan verilere tüm nöronları aktif hale getirmeyen burada negatif değerlere sahip olanları devre dışı bırakan Relu aktivasyon fonksiyonu uygulanmıştır. Sonrasında çok sayıda veri setlerinde kullanımında daha avantajlı olduğu için Adam optimizasyon tekniği kullanılarak 200 iterasyon işlemi gerçekleştirilmiştir.

B. RASTGELE ORMAN (RO)

Breiman tarafından önerilen rastgele orman modeli bir dizi karar ağacına dayanan bir tahmin algoritmasıdır [18]. Yeşilkanat tarafından gerçekleştirilen çalışma da bu modelin hem regresyon hem de sınıflandırma problemlerinde başarılı bir makine öğrenmesi algoritması olduğu söylenmektedir. Ayrıca bu modelin çalışmasında ilk olarak, veri seti öğrenme için eğitim verisi ve öğrenme düzeyini test etmek için doğrulama verisi olmak üzere iki kısma ayrılmaktadır. Ayrıca daha sonrasında bu veri setinden rastgele birçok karar ağacı oluşturulmaktadır. Böylelikle her ağacın büyümesi, düğüm noktalarında rastgele seçilen tahminler tarafından belirlendiği söylenmektedir [19]. Şekil 2'de olduğu gibi rastgele orman algoritmasının kısaca ana çalışma teması incelendiğinde topluluktaki her karar ağacı, girdi özelliklerinin farklı bir rastgele alt kümesine dayalı olarak oluşturulan ve topluluğun son tahminini bu çoğunluk oyuna göre karar veren model olduğu belirtilmiştir. [5], [20].



***Şekil 2**. Rastgele orman yapısı (Random forest structure) [20].*

Bu çalışma kapsamında yapılan incelemeler doğrultusunda RO algoritmasının birden fazla karar ağacı üreterek sınıflandırma işlemini gerçekleştirmektedir. Bu aşamada birbirinden bağımsız olarak çalışan birden fazla karar ağacının bir araya gelerek en iyi sonucun elde edilmesi amaçlanmaktadır. Bu amaç doğrultusunda kullanılacak olan ağaç sayısının artmasıyla ulaşılacak olan nihai sonucun kesinliğinin doğrudan etkileneceği bilindiğinden ağaç sayısı 10, derinlik ise 3 olarak seçilmiştir.

C. K- EN YAKIN KOMŞULUK (KNN)

K-en yakın komşu algoritması hem sınıflandırma hem de regresyon problemlerini çözmek için kullanılan, basit ve uygulaması kolay bir denetimli öğrenme algoritmasıdır. Bu algoritmanın çalışmasındaki temel yapısı modele verilen veri kümesine yeni bir veri eklendiğinde gelen verinin nereye sınıflandırılacağının belirlenmesine dayandığı söylenmektedir [21]. K en yakın komşuluk algoritması için sınıflandırma sorununun çözümlenebilmesi için kullanılan bir algoritma modeli olduğu ve temel veri içerisinde tutulan eğitim seti ile öğrenme işleminin gerçekleştirildiği söylenmektedir. Ayrıca eğitim işlemi gerçekleştirilirken en yakın olduğu tahmin edilen k adet veriyi, belirli uzaklık ölçütü açısından hesaplayarak yaptığı belirtilmektedir [22]. Bu uzaklık hesaplanmasının yapılabilmesi için literatürde genellikle Öklid kullanıldığı ancak bunun yerine Minkowski, Chebyshev ve kosinüs eşitliklerinin kullanabileceği belirtilmiştir [23].

Muja ve arkadaşları tarafından gerçekleştirilen çalışma da KNN algoritması sınıflandırma amacıyla eğitilmiş model içerisine veri gönderildiğinde k tane en yakın temel komşunun etiketlerine bakıldığı ve daha sonrasında sınıf etiketlerinin çoğunluğuna göre ilgili veri kümesine aktarıldığı söylenmektedir [24]. K en yakın komşuluk algoritmasının kullanımının kolay olması, doğrusal olmayan eğitim süresini içermemesi ve gürültülü verilere karşı dayanıklı olması nedeniyle başarılı sonuçlar elde edilmesini sağladığı söylenmektedir. Bunun yanı sıra veri setinin kapasitesine bağlı olarak yüksek bellek gereksinimi duyulan verilerin kullanılmasında işlem yükünün artmasından kaynaklı olarak çalışmasında dezavantajların oluşabileceği söylenebilmektedir [23].

Bu çalışma kapsamında yapılan incelemeler ışığında geliştirilen modelde kullanılan K-NN algoritmasının performansını uzaklık ve kullanılan k değerinin yani komşuluk sayısının etkilediği bilinmektedir. Ayrıca kullanılan k değerindeki artışın karar sınırlarının daha düzgün çalışmasına ama iş yükünün artmasına sebep olabileceği söylenebilmektedir. Bu bilgiler doğrultusunda geliştirilen modelde k değeri 5 olarak seçilirken uzaklık mesafesinin ölçümü için öklid uzaklık hesaplama kriteri kullanılmıştır.

D. ORANGE

Orange Slovenya Ljubljana Üniversitesinde yapay zekâ araştırma ekibi tarafından geliştirilen çekirdeğinde C++ bulunan aynı zamanda veri görselleştirme, veri madenciliği, makine öğrenimi, veri analizi için kullanılabilen açık kaynak kodlu görsel programlama yazılım paketi olarak tanımlanabilmektedir [25], [26]. Orange resmi sitesinde, veri görselleştirme aşamasında basit bir şekilde veri analizi sağlanabildiği, bunun yanı sıra istatistiksel dağılımları, grafikleri, karar ağaçlarını, kümeleme mekanizmasını kullanarak çok boyutlu verileri bile analiz edebileceği söylenmektedir [27].

III. BULGULAR

Bu çalışma kapsamında yapılan incelemelerde protein işlevi işlev tahmini, post-genomik çağdaki en zorlu problemlerden biri olduğu görülmektedir. Yeni tanımlanan proteinlerin sayısı, yüksek verimli tekniklerin gelişmesiyle kümülatif bir şekilde artarak devam etmektedir. Bu durumda literatürde bu problemi çözebilmek adına, özellikle makine öğrenmesi teknolojisinin hızla ilerlemesi ile birden fazla yöntem geliştirilmiştir. Bu durumla birlikte geliştirilen yöntemlerin entegresiyle birçok makine öğrenmesi algoritmasını içinde barındıran araçlar da geliştirilmiştir.

Bu çalışma kapsamında da protein verilerinin artışı ve üzerlerinde yapılan işlemlerin zorlu olması problemi referans alındığı gibi aynı zamanda makine öğrenmesi algoritmalarını içinde barındıran Orange yazılım paketinin kullanımıyla birlikte algoritmalar arasında performans değerlendirilmesi yapılmıştır. Bu değerlendirmenin gerçekleştirilebilmesi için, Gen Ontolojisi terimleri ve Cai ve arkadaşları tarafından [2] yapılan çalışma da paylaşılan veri setindeki insan veri seti kullanılmıştır. Burada sinir ağı, rastgele orman, k-en yakın komşuluk modelleri kullanılarak performans

değerlendirilmesi gerçekleştirilmiştir. Bu değerlendirme işleminin yapılması için kullanılan ORANGE aracı PC platformunda Intel Core i7-9750H CPU 4.50 GHz, 12 MB ön bellek, 16 GB RAM, 1TB HDD, 256 GB SSD ve 6 adet işlemci çekirdeği özelliklerine sahip cihazda çalıştırılarak test işlemleri gerçekleştirilmiştir.

Bu değerlendirme aşamasında kullanılan veri seti öncelikle Orange içinde yer alan sinir ağı modeli kullanılarak eğitim işlemi ve test işlemi gerçekleştirilmiştir. Bunun için öncelikle insan protein verileri ve GO terimlerini içeren veri seti eğitim için hazırlanmıştır. Daha sonrasında ayrıştırılan Mİ veri seti test verisi olarak modele verilmiştir. Bir sonraki aşamada insan veri seti ve insan Mİ veri seti birleştirilerek tahmin aşamasına geçiş gerçekleştirilmiştir. Aynı işlemler BS ve HB veri içeriklerine uygulanmıştır. İnsan veri seti olan ve içinde GO terimlerini içeren veri seti sinir ağları, rastgele orman ve k-en yakın komşuluk kullanılarak eğitilmiştir ve Mİ, BS, HB veri setleri ile test işlemleri gerçekleştirilmiştir. Daha sonrasında set içerisinde çok fazla farklılık olduğu için, alt bölümlemeler çok fazla olmadığından 5 katmanlı çapraz doğrulama yöntemi kullanılmıştır. Daha sonrasında Tablo 2'de yer alan sonuçlar elde edilmiştir.

Tablo 2. *Geliştirilen model testlerinin doğruluk oranları*

Veri Seti İçeriği	**SA**	**RO**	**KNN**
Moleküler İşlev	%88.6	%81.7	%89.0
Biyolojik Süreç	%89.1	%82.6	%89.8
Hücresel Bileşen	%88.2	%81.4	%88.5

Yapılan test işlemlerinin ardından elde edilen sonuçlar Tablo 2'de incelendiğinde değerlerin yaklaşık olduğu ancak k- en yakın komşuluk algoritmasının diğer iki algoritmaya nazaran daha iyi sonuç sergilediği görülebilmektedir.

IV. TARTIŞMA VE SONUÇ

Cai ve arkadaşları tarafından gerçekleştirilen çalışma da bir derin öğrenme modeli tasarlanmıştır. Tasarlanan bu derin öğrenme modelin de hem maya için hem de insan için protein fonksiyon tahmin işlemi gerçekleştirilmiştir. İlgili çalışma da bu tahmin işleminin gerçekleştirilebilmesi için UniProt veri kümesinden veri setleri elde edilmiştir. Daha sonrasında bu veri setleri üzerinde 3-gram ve ProtVec yöntemleri kullanılarak sayısallaştırma işlemi yapılarak protein tabanlı veriler birleştirilerek vektör haline getirilmiştir. Bu birleşimin ardından derin öğrenme modeli kullanılarak tahmin işlemi gerçekleştirilmiştir. Bu tahmin işleminin gerçekleştirilmesinde GO terimleri referans alınmıştır. Bu işlem doğrultusunda moleküler süreç, biyolojik süreç, hücresel bileşen için ayrı ayrı tahmin işlemi gerçekleştirilmiştir. Daha sonrasında 5 katlı çapraz doğrulama işlemi kullanılmış ve nihai sonuç elde edilmiştir. Elde edilen sonuçlardan başarı oranları incelendiğinde moleküler işlev için %95, biyolojik süreç için %92, hücresel bileşen için %95 oranında veriler elde edildiği söylenmektedir [2].

Bu çalışma kapsamında Cai ve arkadaşları [2] tarafından paylaşılan veri seti kullanılarak geçmiş çalışma da kullanılan CNN modeli yerine Orange aracının sağladığı sinir ağı, rastgele orman ve k-en yakın komşuluk algoritmalarının kullanıldığı modeller ayrı ayrı Orange platformu üzerinde kullanılarak tasarlanmıştır. Yapılan tasarımların kullanımıyla elde edilen sonuçlar incelendiğinde sinir ağı modeli kullanılarak yapılan işlemlerde moleküler işlev için %88.6, biyolojik süreç için %89.1, hücresel bileşen için %88.2 başarı oranı elde edilmiştir. Rastgele orman modeli kullanılarak yapılan işlemlerde moleküler işlev için %81.7, biyolojik süreç için %82.6, hücresel bileşen için %81.4 başarı oranı elde edilmiştir. K-en yakın komşuluk modeli kullanılarak yapılan işlemlerde moleküler işlev için %89.0, biyolojik süreç için %89.8, hücresel bileşen için %88.5 başarı oranı elde edilmiştir. Yapılan çalışma sonucunda sinir ağı, rastgele orman ve k-en yakın komşuluk modelleri kendi içinde incelendiğinde en iyi sonucun k-yakın komşuluk modeli kullanılarak elde edildiği en kötü sonucun ise rastgele orman ile elde edildiği tespit edilmiştir.

Elde edilen sonuçlar incelendiğinde literatürde yer alan diğer çalışmalardaki (Tablo 1) başarılarla kıyaslandığında daha iyi sonuçlarında elde edildiği göründüğü gibi aynı zamanda benzer sonuçlarında elde edildiği görülmüştür. Bu çalışma kapsamında asıl amacın Orange aracı olduğu göz önüne alındığında kullanılan veri setinin büyüklüğü doğrultusunda başarı oranının kabul edilebilir düzeyde olduğu söylenebilmektedir. Ancak gerçekleştirilen test aşamalarında kullanılan cihazın özellikleri de dikkate alındığında sürenin normal şartlara göre daha uzun sürdüğü ve kasmaların olduğu gözlemlenmiştir. Dolayısıyla bu bilgi ışığında kullanılacak olan veri setinin boyutu dikkate alındığında kolaylık açısından Orange aracının kullanımının avantaj sağlayabileceği gibi veri miktarının çok artması söz konusu olduğunda ortaya bir dezavantajın çıkabileceği söylenebilmektedir.

Son olarak gelecek çalışmalarda Orange aracı üzerinde uygun veri setleri kullanılarak ve aracın sağladığı algoritmaların kullanımıyla model tasarlanarak hayati öneme sahip olan protein fonksiyonlarının tahmin işleminin gerçekleştirilebileceği öngörülebilmektedir.

V. KAYNAKLAR

[1] https://tr.wikipedia.org/wiki/Protein (Erişim Tarihi: 30.04.2022)

[2] Y. Cai, J. Wang, ve L. Deng, "Sdn2go: An İntegrated Deep Learning Model For Protein Function Prediction", *Front. Bioeng. Biotechnol.*, c. 8, Sayı April, Ss. 1–11, 2020

[3] J. R. Hoffman Ve M. J. Falvo, "Protein- Which İs Best?", *J. Sport. Sci. Med.*, c. 3, Sayı 3, Ss. 118–130, 2004.

[4] İ. Alakuş, Talha Burak; Türkoğlu, "İnsana Ait Protein Fonksiyonlarının Protein Haritalama Teknikleri ve Derin Öğrenme Modeli ile Tahmin Edilmesi Prediction Of Human Protein Functions W", *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi Pamukkale University Journal Of Engineering Sciences* c. 28, Sayı X, Ss. 1–12, 2021

[5] K. Hakala vd., "Neural Network And Random Forest Models İn Protein Function Prediction", *IEEE/Acm Transactions On Computational Biology And Bioinformatics.* 2020

[6] S. Pitre Vd., "Pıpe: A Protein-Protein İnteraction Prediction Engine Based On The Re-Occurring Short Polypeptide Sequences Between Known İnteracting Protein Pairs", *Bmc Bioinformatics*, c. 7, Ss. 1–15, 2006

[7] N. Fukuhara ve T. Kawabata, "Homcos: A Server To Predict İnteracting Protein Pairs And İnteracting Sites By Homology Modeling Of Complex Structures.", *Nucleic Acids Res.*, c. 36, Sayı Web Server İssue, Ss. 185–189, 2008

[8] J. Shen Vd., "Predicting Protein-Protein İnteractions Based Only On Sequences Information", *Proc. Natl. Acad. Sci. U. S. A.*, c. 104, Sayı 11, Ss. 4337–4341, 2007

[9] L. Cai, Z. Pei, S. Qin, ve X. Zhao, "Prediction Of Protein-Protein İnteractions İn Saccharomyces Cerevisiae Based On Protein Secondary Structure", *Proc. - 2012 Int. Conf. Biomed. Eng. Biotechnol.* İcbeb 2012, Ss. 413–416, 2012

[10] M. Yeni, B. Bilim, M. Polat, ve A. G. Karahan, "Multidisipliner Yeni Bir Bilim Dalı: Biyoinformatik Ve Tıpta Uygulamaları", *Sdü Tıp Fakültesi Dergisi.*, c. 16, Sayı 3, Ss. 41–50, 2009.

[11] İ. Kösesoy, "Konak-Patojen Protein Etkileşiminin Hesaplamalı Yöntemler İle Tahmini", 2018.

[12] “Biyoinformatikte Makine Öğrenmesi ve Teknikleri – Pharmaino Science”. *https://Pharmaino.Com/Biyoinformatikte-Makine-Ogrenmesi-Ve-Teknikleri/* (Erişim May. 20, 2022).

[13] M. Kulmanov ve R. Hoehndorf, “Deepgoplus: Improved Protein Function Prediction From Sequence”, *Bioinformatics*, c. 36, Sayı 2, Ss. 422–429, 2020

[14] B. A. Sokhansanj ve G. L. Rosen, “Mapping Data To Deep Understanding: Making The Most Of The Deluge Of Sars-Cov-2 Genome Sequences”, *Msystems*, Sayı February, 2022

[15] S. Gelman, S. A. Fahlberg, P. Heinzelman, P. A. Romero, ve A. Gitter, “Neural Networks To Learn Protein Sequence-Function Relationships From Deep Mutational Scanning Data”, *Proc. Natl. Acad. Sci. U. S. A.*, c. 118, Sayı 48, 2021

[16] E.Atar, “Yapay Sinir Ağları ile Proteinlerin İkincil Yapılarının Kestirimi”, Yüksek Lisans Tezi, Elektronik ve Haberleşme Mühendisliği, Yıldız Teknik Üniversitesi, İstanbul, Türkiye 2005.

[17] https://en.wikipedia.org/wiki/Neural_network (Erişim Tarihi: 10.05.2022)

[18] L. Breıman, “Random Forest”, *Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, c. 12343 Lncs, Ss. 503–515, 2020

[19] C. M. Yeşilkanat, “Spatio-Temporal Estimation Of The Daily Cases Of Covıd-19 In Worldwide Using Random Forest Machine Learning Algorithm”, *Chaos, Solitons And Fractals*, c. 140, 2020

[20] C. Nguyen, Y. Wang, ve H. N. Nguyen, “*Random Forest Classifier Combined With Feature Selection For Breast Cancer Diagnosis And Prognostic*”, c. 2013, Sayı May, Ss. 551–560, 2013.

[21] K. Özdemir, “K-En Yakın Komşu Algoritması (K-Nearest Neighbor Algorithm) https://Medium.Com/Batech/K-En-Yakın-Komşu-Algoritması-K-Nearest-Neighbors-Algorithm-16e5ab69af2e. (Erişim Tarihi: 10.05.2022)

[22] S. A. Dudani, “*The Distance-Weighted K-Nearest-Neighbor Rule*”, Ss. 325–327.

[23] M. A. Pala, M. E. Çimen, Ö. F. Boyraz, M. Z. Yildiz, ve A. F. Boz, “Meme Kanserinin Teşhis Edilmesinde Karar Ağacı Ve Knn Algoritmalarının Karşılaştırmalı Başarım Analizi”, *Acad. Perspect. Procedia*, c. 2, Sayı 3, Ss. 544–552, 2019

[24] M. Muja Ve D. G. Lowe, “Fast Approximate Nearest Neighbors With Automatic Algorithm Configuration”, *Vısapp 2009- Proc. 4th Int. Conf. Comput. Vis. Theory Appl.*, c. 1, Ss. 331–340, 2009,

[25] https://en.wikipedia.org/wiki/Orange_(software) (Erişim Tarihi: 01.06.2022)

[26] M. Kaya Keleş ve S. Özel, “Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Karşılaştırılması”, *Akad. Bilişim’14 - Xvı. Akad. Bilişim Konf. Bildir.*, Ss. 47–53, 2014.

[27] https://orangedatamining.com/ (Erişim Tarihi: 01.06.2022)

[28] Çakmak E, Selvi İ. "Derin Öğrenme (CNN, RNN, LSTM, GRU) Kullanarak Protein İkincil Yapı Tahmini". *Acta Infologica* 2022;0:0–0

[29] Aydin Z, Kaynar O, Görmez Y, Işik YE. "Comparison of machine learning classifiers for protein secondary structure prediction". *26th IEEE Signal Process Commun Appl Conf* SIU 2018 2018:1–4.

[30] Li YH, Xu JY, Tao L, Li XF, Li S, Zeng X, vd. SVM-prot 2016: "A web-server for machine learning prediction of protein functional families from sequence irrespective of similarity". *PLoS One* 2016;11:1–14.

[31] Bonetta R, Valentino G. "Machine learning techniques for protein function prediction". *Proteins Struct Funct Bioinforma* 2020;88:397–413

[32] Sureyya Rifaioglu A, Doğan T, Jesus Martin M, Cetin-Atalay R, Atalay V. DEEPred: "Automated Protein Function Prediction with Multi-task Feed-forward Deep Neural Networks". *Sci Rep* 2019;9:1–16.