Pamukkale Univ Muh Bilim Derg, 28(2), 255-265, 2022

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

İnsana ait protein fonksiyonlarının protein haritalama teknikleri ve derin öğrenme modeli ile tahmin edilmesi

Prediction of human protein functions with protein mapping techniques and deep learning model

Talha Burak ALAKUŞ[1*], *İbrahim TÜRKOĞLU*[2]

[1]Yazılım Mühendisliği Bölümü, Mühendislik Fakültesi, Kırklareli Üniversitesi, Kırklareli, Türkiye.
talhaburakalakus@klu.edu.tr

[2]Yazılım Mühendisliği Bölümü, Teknoloji Fakültesi, Fırat Üniversitesi, Elazığ, Türkiye.
iturkoglu@firat.edu.tr

Geliş Tarihi/Received: 11.01.2021 Düzeltme Tarihi/Revision: 15.03.2021 doi: 10.5505/pajes.2021.51261
Kabul Tarihi/Accepted: 22.03.2021 Araştırma Makalesi/Research Article

Öz

Canlıların moleküler mekanizmasının anlaşılabilmesi için protein fonksiyonları önem arz etmektedir. Proteinlere ait fonksiyonlar belirlenirken, proteinlerin yapılarından yararlanılır. Protein fonksiyonları daha çok, karakterize edilmemiş protein dizilimlerinin anotasyonlarının belirleyebilmek, canlıların hücresel mekanizmalarını anlayabilmek, genlerde ya da proteinlerde hastalığa neden olan fonksiyonel değişiklikleri belirleyebilmek ve hastalıkların önlenebilmesi, tedavi edilebilmesi ve teşhisi için yeni yaklaşımlar geliştirmek için kullanılmaktadır. Protein fonksiyonları deneysel yöntemlerle etkin bir şekilde belirlenebilmektedir. Ancak, deneysel yöntemlerin zaman alması ve çok sayıda kimyasal süreçten geçmesi, bu aşamaların yavaş ve maliyetli olmasına neden olmaktadır. Bunlara ek olarak, fonksiyonel yapısı ve dizilimi bilinen bazı proteinlerin anotasyonları deneysel süreçlerden dolayı halen belirlenememektedir. Bu gibi nedenler ve dezavantajlardan dolayı hesaplama-tabanlı uygulamalara ihtiyaç duyulmaktadır. Hesaplama-tabanlı uygulamalar için genellikle yapay zeka algoritmaları kullanılmaktadır. Yapay zeka yöntemleri ile protein fonksiyonlarının tahmin edilebilmesi için protein dizilimlerinin belirli haritalama yöntemleri ile sayısal hale getirilmesi gerekmektedir. Bu çalışmada, belirli protein haritalama teknikleri kullanılarak gen ontoloji tabanlı protein fonksiyonlarının tahmini gerçekleştirilmiştir. Çalışma, protein verilerinin elde edilmesi, protein dizilimlerinin sayısallaştırılması, protein fonksiyonlarının sınıflandırılması ve protein haritalama tekniklerinin performanslılarının belirlenmesi olmak üzere dört farklı aşamadan oluşmaktadır. Çalışmanın sonunda, biyolojik süreç kategorisinde en iyi doğruluk ve AUC skoru PAM250 protein haritalama tekniği ile elde edilmiş ve sırasıyla %69 ve %88 olarak hesaplanmıştır. Hücresel bileşen kategorisinde ise en iyi doğruluk ve AUC değer, sırasıyla %64 ve %89 oranı ile FIBHASH protein haritalama tekniği ile elde edilmiştir. Moleküler fonksiyon kategorisinde ise %64 AUC oranı ve %89 doğruluk değeri ile en iyi sonuç FIBHASH ile elde edilmiştir. Önerilen yapay zekâ yöntemi ile protein sayısal haritalama tekniklerinin birlikte kullanımının, protein fonksiyonlarının tahmin edilmesinde etken bir role sahip olduğu gözlemlenmiştir.

Anahtar kelimeler: Protein fonksiyonları, Derin öğrenme, Protein haritalama teknikleri, İki yönlü uzun-kısa vadeli bellek.

Abstract

Protein functions are important for understanding the molecular mechanism of living organisms. Protein structures are used when determining the functions of proteins. Protein functions are mostly used to determine the annotations of uncharacterized protein sequences, to understand the cellular mechanisms of living things, to identify functional changes in genes or proteins that cause disease, and to develop new approaches to prevent, treat and diagnose diseases. Protein functions can be determined effectively by experimental methods. However, experimental methods take time and go through many chemical processes, causing these stages to be slow and costly. In addition to these, the annotations of some proteins whose functional structure and sequence are known cannot be specified due to experimental processes. Due to such reasons and disadvantages, computational-based approaches are needed. Artificial intelligence algorithms are generally used for computational-based applications. In order to predict protein functions with artificial intelligence methods, protein sequences must be mapped with certain mapping methods. In this study, prediction of gene ontology-based protein functions was performed using certain protein mapping techniques. The study consists of four different stages; obtaining protein data, mapping protein sequences, classifying protein functions, and determining the performance of protein mapping techniques. At the end of the study, the best accuracy and AUC score in the biological process category was obtained by the PAM250 protein mapping technique and was calculated as 69% and 88%, respectively. In the cellular component category, the best accuracy and AUC value were obtained by FIBHASH protein mapping technique with 64% and 89%, respectively. In the molecular function category, the best result was obtained with FIBHASH with 64% AUC score and 89% accuracy. It has been observed that the combined use of the proposed artificial intelligence method and protein numerical mapping techniques have an effective role in predicting protein functions.

Keywords: Protein functions, Deep learning, Protein mapping techniques, Bidirectional long-short term memory.

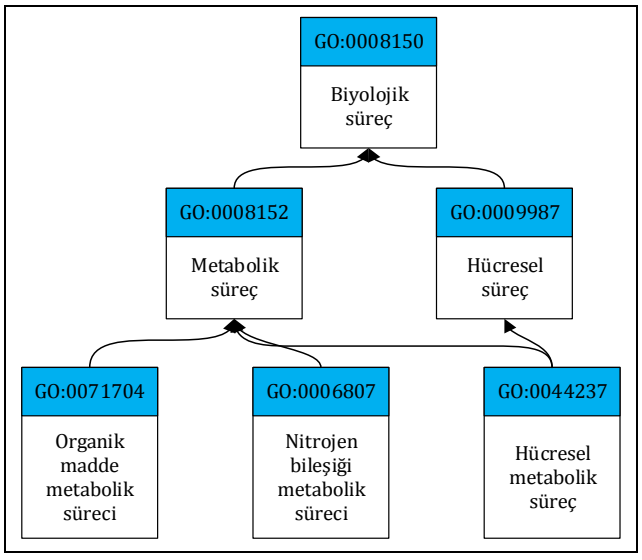
1 Giriş

Doğada 20 çeşit amino asit bulunmakta ve bu amino asitler DNA (Deoksiribo Nükleik Asit) dizilimleri tarafından kodlanmaktadır. Proteinler, hücre sinyalizasyonu, regülasyon, reaksiyonların katalizlenmesi, membran transportu, amino asit yapısının sağlanması başta olmak üzere hücrelerde önemli roller üstlenmektedir. Bir proteinin fonksiyonel işlevi onun yapısına bağlı olmaktadır. Proteinlerin fonksiyonel anotasyonlarının (dizilimlerinin) belirlenmesi, hücresel aktivitelerin mekanizmalarını anlamak, proteinlerde ya da genlerde hastalığa neden olan fonksiyonel değişiklikleri

*Yazışılan yazar/Corresponding author

tanımlamak ve hastalıkların önlenmesi, teşhisi ve tedavisi için büyük önem arz etmektedir [1]. Protein fonksiyonları, genellikle in-vitro (laboratuvar ortamı) ve in-vivo (canlı ortam) yöntemleri ile elde edilmektedir. Ancak bu tarz işlemlerde kimyasal sürecin uzun olması hem zaman açısından hem de iş yükü açısından maliyetli olmaktadır [2],[3]. Bu nedenden dolayı, protein sekans verilerinin sürekli artması nedeniyle ortaya çıkan bilgi açığı yeterli sürede çözülememektedir [4]. Bunlara ek olarak, yeni proteinlerin anotasyonları sürecin yavaş işlemesinden dolayı hızlı bir şekilde belirlenememektedir [5]. Bu dezavantajlardan dolayı, proteinlerin anotasyonlarını otomatik bir şekilde belirleyebilmek ve protein fonksiyonlarını tahmin edebilmek için hesaplama-tabanlı yöntemler sıklıkla kullanılmaktadır.

Protein fonksiyonları, biyokimyasal, fizyolojik, fenotipik ve hücresel olmak üzere farklı karmaşıklık seviyelerine göre tanımlanmaktadır. Bunlara ek olarak, protein fonksiyonları hiyerarşik bir şekilde de tanımlanabilmektedir. Hesaplama-tabanlı çalışmalarda, en sık hiyerarşik yaklaşımlar değerlendirilmekte ve bunun için ise GO (Gen Ontolojisi) terimleri kullanılmaktadır. GO terimleri, protein fonksiyonlarının farklı seviyelerini ifade etmektedir [6]. Bir proteinin moleküler ya da biyokimyasal fonksiyonunu, dizilim ya da yapısal bilgilere dayanarak elde etmek büyük önem arz etmektedir. Bundan dolayı in-silico (benzetim-tabanlı/hesaplama-tabanlı) yaklaşımlar protein fonksiyonlarının tahmininde yardımcı olabilmektedir [7]. MF (Molecular Function - Moleküler Fonksiyon), BP (Biological Process - Biyolojik Süreç) ve CC (Cellular Component - Hücresel Bileşen) olmak üzere, birbirine bağımlı seviyelerden oluşan üç farklı GO terimi bulunmaktadır [8]. Her bir GO terimi, eşsiz bir fonksiyonel niteliği ifade eder ve tüm GO terimleri, yönlendirilmiş döngüsel olmayan grafik yapısıyla birebiriyle ilişkilendirilir. Şekil 1'de belirli GO terimlerine ait hiyerarşik yapı verilmiştir.



Şekil 1. Bazı GO terimlerine ait yönlendirilmiş döngüsel olmayan grafik yapısı.

Figure 1. Directed acyclic graph structure of some GO terms.

Moleküler fonksiyon, moleküler düzeydeki aktiviteyi ifade etmektedir. Biyolojik süreç, moleküler fonksiyonların bir araya gelmesiyle gerçekleştirilen daha geniş çaplı fonksiyonlar olarak tanımlanmaktadır. Son olarak ise hücresel bileşen, proteinin hücre içinde fonksiyonunu gerçekleştirdiği yer ya da yerler olarak ifade edilmektedir.

Teknolojik gelişmeler protein dizilimlerinin elde edilmesini kolaylaştırmış, ancak buna bağlı olarak ise protein anotasyonlarının belirlenmesini zorlaştırmıştır. Çünkü protein anotasyonlarının belirlenebilmesi için deneysel yöntemlere ihtiyaç duyulmakta, bu durum da sürecin yavaş ilerlemesine neden olmaktadır. Bunlara ek olarak, anotasyon işlemleri manuel olarak yapılmaktadır ve bundan dolayı protein dizilimlerinin sayısı ve fonksiyonel anotasyonları arasındaki fark sürekli artış göstermektedir [5]. Bu nedenden dolayı, biyoinformatik alanında karşılaşılan en büyük zorluklardan bir tanesi, proteinlerin biyolojik süreçlerde ve hastalıklarda oynadıkları rolü ve protein fonksiyonlarının gerçekleştirdiği mekanizmayı tahmin etmektir [5]. Bu problemleri çözebilmek için çok sayıda yeni algoritmalar geliştirilmekte, ancak bu algoritmaların hemen hemen hepsi geleneksel yaklaşımlara yani manuel işlemlere dayanmaktadır. Bundan dolayı bu aradaki boşluğu kapatabilmek için geleneksel yöntemlerden ziyade hesaplama-tabanlı yöntemlerin ihtiyacı artmıştır.

Protein fonksiyonlarının hesaplama-tabanlı yöntemler ile tahmin edilebilmesi için yapay öğrenme yöntemleri kullanılmaktadır. Bu amaçla, makine öğrenmesi ve derin öğrenme yöntemleri sıklıkla bu alana uygulanmaktadır. Araştırmacılar, dizilim-fonksiyon ilişkilerini belirleyebilmek için sıklıkla makine öğrenmesi algoritmalarına başvurmaktadırlar. Protein fonksiyonlarının tam olarak anlaşılmaması durumunda bile, makine öğrenmesi yöntemlerinin bu alanda etkili ve başarılı olduğu gözlemlenmiştir [9]. Ancak, bazı araştırmalar, makine öğrenmesi algoritmalarının, proteinlerin fonksiyon bilgisinin net olduğu durumlarda etkili olduğunu, aksi takdirde yanlış hizalama sonuçları üretebildiğini göstermiştir [10].

Bunlara ek olarak, makine öğrenmesi algoritmalarının karşılaştığı en büyük problemlerden birisi de sınıflandırma performansındaki düşüştür [10]. Protein fonksiyonlarının çok sınıflı bir yapıya sahip olması ve biyolojik rollerin hiyerarşik GO terimleri ile ifade edilmesi, makine öğrenmesi algoritmalarının karşılaştığı en büyük sorunlardan birisidir [11]. Makine öğrenmesi algoritmalarında yaşanan problemlerden dolayı, derin öğrenme algoritmalarının bu alanda da etkili bir şekilde kullanıldığı gözlemlenmiştir. Derin öğrenmenin büyük veri alanında başarılı olması ve makine öğrenmesinin sınıflandırma ve regresyon problemlerinde karşılaştığı sorunlarla karşılaşmaması, derin öğrenme algoritmalarını bu alanda popüler kılmıştır [12],[13]. Protein fonksiyonlarının yapay öğrenme yöntemleri ile tahmin edilebilmesi için, protein dizilimlerinin sayısallaştırılması gerekmektedir. Bilindiği üzere, protein dizilimleri amino asitlerden oluşmakta ve her bir aminoaside karşılık gelen bir kod bulunmaktadır. Bu kodlar ise harflerden meydana gelmektedir. Protein dizilimleri üzerinde bir ön işlem yapmadan (sayısallaştırmadan), protein fonksiyonlarını yapay öğrenme teknikleri ile tahmin etmek mümkün olmamaktadır.

Bu çalışmada, farklı protein haritalama yöntemleri kullanılarak, insana ait protein fonksiyonlarının tahmin işlemi gerçekleştirilmiş ve bu doğrultuda BiLSTM (Bidirectional Long Short Term Memory-İki Yönlü Uzun Kısa Süreli Bellek) kullanılmıştır. Çalışmanın ana amacı literatürde etkili ve başarılı olan protein haritalama tekniklerini, toplamış olduğumuz veri kümesi üzerinde kullanarak performanslarını kıyaslamaktır. Çalışma protein verilerinin elde edilmesi,

protein dizilimlerinin sayısallaştırılması, sınıflandırma işleminin gerçekleştirilmesi ve protein haritalama tekniklerinin başarım ölçütlerinin değerlendirilmesi olmak üzere dört aşamadan meydana gelmektedir. Birinci aşamada üç farklı GO kategorisine ait protein dizilimleri UniProt veri kümesinden elde edilmiştir. Ardından ikinci aşamada, elde edilen protein dizilimleri beş farklı protein haritalama yöntemi ile sayısallaştırılmıştır. Sayısallaştırma işlemi için EIIP (Electron-Ion Interaction Potential-Elektron-İyon Etkileşim Potansiyeli), Atchley faktörleri, FIBHASH (FibonacciHash), PAM250 (Point Accepted Mutation-Nokta Kabul Edilen Mutasyon) ve BLOSUM62 (Blocks Substitution Matrix - Blok Değiştirme Matrisi) yöntemleri kullanılmıştır.

Sayısallaştırma işleminin ardından üçüncü kısımda BiLSTM modeli tasarlanmış ve sınıflandırma işlemi için kullanılmıştır. Son aşamada ise protein haritalama tekniklerinin başarımları doğruluk, kesinlik, hassasiyet, en yüksek f skor ve AUC (Area Under Curve- Eğri Altında Kalan Alan) skorları ile belirlenmiştir. Çalışmanın sonunda, BP kategorisinde en iyi doğruluk ve AUC skoru PAM250 protein haritalama tekniği ile elde edilmiş ve sırasıyla %69 ve %88 olarak hesaplanmıştır. CC kategorisinde ise en iyi doğruluk ve AUC değer, sırasıyla %64 ve %89 oranı ile FIBHASH protein haritalama tekniği ile elde edilmiştir. MF kategorisinde ise %89 AUC oranı ve %64 doğruluk değeri ile en iyi sonuç FIBHASH ile elde edilmiştir. Çalışmanın öne çıkan katkıları şu şekilde özetlenebilir;

- Çalışmada protein haritalama tekniklerinin performansları değerlendirilmiş ve kullanılan protein haritalama tekniklerinin tahmin işleminde önemli bir faktör olduğu gözlemlenmiştir,
- Sadece dizilim-tabanlı bir yaklaşımın protein fonksiyonlarını tahmin etmede etkili olduğu gözlemlenmiştir,
- Alanyazın araştırıldığında bu alanda Türkçe yazılmış bir makaleye rastlanılmamıştır. Bu anlamda bu çalışma, Türkçe alanyazına katkı sağlayacaktır.

Çalışmanın geri kalan kısmı şu şekilde organize edilmiştir. İkinci kısımda, protein fonksiyonlarının tahmini için gerçekleştirilmiş olan hem derin öğrenme hem de makine öğrenmesi tabanlı çalışmalar irdelenmiştir. Üçüncü kısımda kullanılan veri kümesi hakkında bilgiler verilmiştir. Ayrıca, bu kısımda, kullanılan protein haritalama teknikleri incelenmiş ve amino asitlerin sayısal karşılıkları verilmiştir. Dördüncü kısımda geliştirilmiş olan derin öğrenme algoritması hakkında bilgi verilmiş ve parametrelerinden bahsedilmiştir. Bunlara ek olarak, bu kısımda protein haritalama tekniklerinin sınıflandırma başarımları hesaplanmıştır. Beşinci kısımda elde edilen sonuçlar tartışılmıştır. Son kısımda ise çalışmanın gelecekteki kullanım alanları üzerinde durulmuş ve çalışmanın alanyazına katkısı sunulmuştur.

2 İlgili çalışmalar

Bu kısımda protein fonksiyonun tahminine yönelik gerçekleştirilen hesaplama-tabanlı çalışmalardan bahsedilmiştir. Çalışmalarda kullanılan haritalama yöntemleri, derin öğrenme ve makine öğrenmesi algoritmaları ve kullanılan veri kümeleri hakkında bilgiler verilmiştir. [14] No.lu çalışmada araştırmacılar, yeni bir derin öğrenme modeli geliştirerek protein fonksiyonlarını tahmin etmişlerdir. Çalışmada kullanılan protein verileri UniProt veri kümesinden elde edilmiş ve insan ve maya olmak üzere iki farklı türe ait proteinler değerlendirilmiştir. Protein fonksiyonlarını tahmin edebilmek için hem insan hem de maya türüne ait dizilim tabanlı, protein-protein etkileşim tabanlı ve protein bölgesi tabanlı veriler kullanılmıştır. Çalışmada protein dizilimleri hem 3-gram hem de ProtVec yöntemleri kullanılarak sayısallaştırılmıştır. Sayısallaştırma işleminin ardından, dizilim tabanlı, protein-protein etkileşim tabanlı ve protein bölgesi tabanlı veriler birleştirilmiş ve vektör haline getirilmiştir. Birleştirme işleminin ardından geliştirmiş oldukları CNN (Convolutional Neural Networks-Evrişimsel Sinir Ağları) tabanlı derin öğrenme modelini kullanarak protein fonksiyonlarını tahmin etmişlerdir.

Tahmin işlemi GO terimlerine göre gerçekleştirilmiştir. Bu işlem BP, MF ve CC için ayrı ayrı yapılmıştır. Önerilen derin öğrenme modelinin performansı en yüksek f skor, TRAUC (True Rate Area Under Curve - Doğru Oran Altında Kalan Alan) ve AUC ölçütleri ile belirlenmiştir. Doğrulama işlemi için 5 katlı çapraz-doğrulama yöntemi kullanılmıştır. Çalışmanın sonunda, insana ait proteinler için en yüksek f skorunu %65.3 ile MF için elde etmişlerdir. Bu oran CC için %61.7, BP için ise %50.7 olarak hesaplanmıştır. Maya için ise en yüksek f skoru MF için elde edilmiş ve %61.1 oranında bir değer hesaplanmıştır. En yüksek f skor değeri CC ve BP için sırasıyla %52.8 ve %41.5 olarak hesaplanmıştır. Başka bir çalışmada ise araştırmacılar protein fonksiyonlarını tekrarlayıcı sinir ağları ile tahmin etmişlerdir [15].

Çalışmada protein verileri UniProt veri kümesinden elde edilmiş ve toplamda 523.990 adet protein dizilimi ve 42.918 adet GO terimleri değerlendirilmiştir. CAFA3 veri kümesi kullanılarak toplamda 23 türe ait 130.787 adet protein dizilimi önerilen yöntem ile test edilmiştir. Protein dizilimleri ProLan yöntemi, GO terimleri ise GOLan yöntemi ile sayısallaştırılmıştır. Sayısallaştırma işleminin ardından üç katmanlı bir tekrarlayıcı sinir ağı modeli tasarlanmış ve fonksiyonların tahmin işlemi gerçekleştirilmiştir. Önerilen yöntemin performansı AUC skoru ile belirlenmiş ve sonuç %33.3 olarak elde edilmiştir. [16] No.lu çalışmada araştırmacılar, protein fonksiyonlarını protein dizilim ve etkileşim bilgilerini kullanarak tahmin etmişlerdir. Çalışmada GO terimlerine dayalı bir tahmin işlemi gerçekleştirilmiş ve toplamda 60.710 adet protein dizilimi kullanılırken, 27.760 adet sınıf değerlendirilmiştir. BP kategorisi için 19.181, MF kategorisi için 6.221 ve CC kategorisi için ise 2.358 adet etiket sınıflandırılmıştır. Veri kümesinde kullanılan protein anotasyonlarının %90'ından fazlası deneysel olarak kanıtlanmıştır. Protein dizilimleri tek-vektör kodlama (one-hot encoding) yöntemi ile sayısallaştırılmıştır. Sayısallaştırma işleminin ardından evrişimsel sinir ağı modellenmiş ve GO terimlerinin sınıflandırılması gerçekleştirilmiştir.

Önerilen yöntemin başarımı en yüksek f skor, ortalama hassasiyet, ortalama kesinlik, AUC ve MCC (Matthews Correlation Coefficient - Matthews Korelasyon Katsayısı) ile belirlenmiştir. Test işlemi için CAFA3 veri kümesi kullanılmış ve BP, MF ve CC için ayrı ayrı sonuçlar irdelenmiştir. BP kategorisi için 0.34 oranında en yüksek f skor elde edilmiş ve bu kategori için AUC skoru 0.88 olarak hesaplanmıştır. MF kategorisi için AUC skoru 0.90 olarak hesaplanırken, en yüksek f skor 0.47 olarak belirlenmiştir. Çalışmada en iyi başarım CC kategorisi için elde edilmiş ve sırasıyla en yüksek f skor ve AUC için 0.52 ve 0.95 oranında değerler gözlemlenmiştir. [17] No.lu çalışmada yazarlar, dizilim-tabanlı protein fonksiyonlarını tahmin edebilmek için öğrenilen sırayı öğrenmek (learning to rank) tabanlı yeni bir yöntem geliştirmişlerdir. Çoğu çalışmada

olduğu gibi bu çalışma da GO terimlerine dayalı bir tahmin işlemi gerçekleştirilmiştir.

Protein dizilimleri UniProt veri kümesinden elde edilirken, GO terimleri SwissProt veri kümesinden elde edilmiştir. Protein dizilimleri K-mer, InterProt ve ProFet yöntemleri ile sayısallaştırılmış ve öğrenilen sırayı öğrenmek algoritması ile sınıflandırılmıştır. Önerilen yöntemin başarımı AUC ve en yüksek f skor ile belirlenmiştir. En iyi AUC skoru CC kategorisinde elde edilmiş ve 0.7 olarak hesaplanmıştır. Bu oran MF kategorisi için 0.55 olarak hesaplanırken, BP için 0.23 olarak belirlenmiştir. AUC skorda olduğu gibi en yüksek f skor CC kategorisinde gözlemlenmiş ve 0.69 olarak bulunmuştur. Bu oran MF kategorisinde 0.58 ve BP kategorisinde ise 0.37 olarak hesaplanmıştır. [18] No.lu çalışmada araştırmacılar sinir ağı ve rastgele orman modellerini kullanarak protein fonksiyonlarını tahmin etmişlerdir. Çalışmada kullanılan eğitim verileri SwisProt veri kümesinden elde edilmiştir. Eğitim veri kümesinde toplamda 67.118 adet protein için 387.416 adet anotasyon kullanılmıştır. Protein dizilimleri BLAST özellikleri, InterproScan, taksonomi özellikleri, dizilim özellikleri ve amino asit indeksi olmak üzere çeşitli yöntemler ile sayısallaştırılmıştır. Ardından rastgele orman, ileri beslemeli sinir ağı ve evrişimsel sinir ağı modellenmiş ve bu modeller birleştirilerek sınıflandırma işlemi yapılmıştır. Modelin başarımı f skor, hassasiyet ve kesinlik değerleri ile ölçülmüştür. Çalışmanın sonunda 0.49 oranında f skor, 0.45 oranında kesinlik ve 0.55 oranında hassasiyet değerleri elde edilmiştir. [58] No.lu çalışmada araştırmacılar graf tabanlı bir yaklaşım önererek proteinlerin fonksiyonlarını tahmin etmişlerdir.

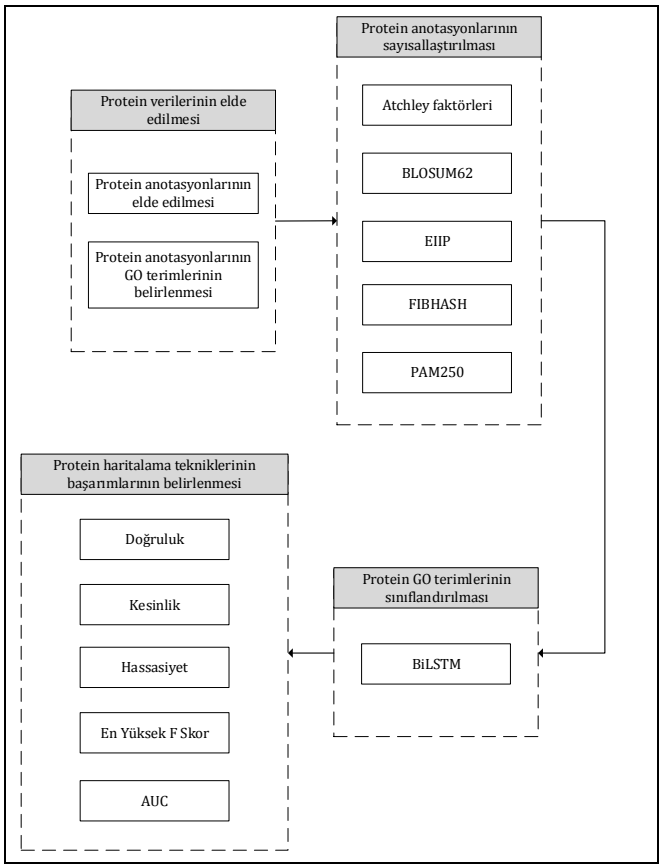
Çalışmada dizilim-tabanlı bir yaklaşım kullanılmış ve CC, BP ve MF GO terimlerine dayalı sınıflandırma yapılmıştır. Zebra balığı, insan, fare, Avustralya kurbağası ve meyve sineğine olmak üzere beş farklı türe ait protein dizilimleri kullanılmıştır. Protein dizilimleri proteinlerin benzerlik oranına bakılarak sayısallaştırılmıştır. Benzerlikleri belirlemek için graf teorisinden yararlanılmıştır. Çalışmanın sonunda CC kategorisi için ortalama %79.27, MF için %71.63 ve BP için % 55.80 oranında bir doğruluk elde edilmiştir. Başka bir çalışmada ise araştırmacılar protein fonksiyonlarının tahmin edem sistem geliştirmeyi hedeflemişlerdir [59].

Farekulağı teresine ait veriler kullanılmış ve bu amaç için sadece BP GO terimi göz önünde bulundurulmuştur. Protein dizilimleri çeşitli benzerlik algoritmaları ile sayısallaştırılmış ve sınıflandırma işlemi için ise kNN (k Nearest Neighbors-k En Yakın Komşular) algoritması kullanılmıştır. Sınıflandırıcının performansı AUC değerlendirme ölçütü ile hesaplanmış ve %77.00 oranında bir başarım elde edilmiştir. [60] No.lu çalışmada yazarlar derin öğrenme tabanlı bir sistem tasarlayarak proteinlerin fonksiyonlarını tahmin etmişlerdir. Çalışmada rastgele yürüyüş (random walk) tabanlı ağ modeli geliştirilmiştir. Protein dizilimleri maya ve insandan elde edilmiştir. Geliştirilen sınıflandırıcının performansı f1 skoru ile hesaplanmıştır. Maya türü için en yüksek f1 skor MF GO terimi için elde edilmiş ve yaklaşık %27 oranında bir hesaplama gerçekleştirilmiştir. İnsan için ise en yüksek oran gene MF kategorisinden elde edilmiş ve f1 skor yaklaşık olarak %18 olacak şekilde hesaplanmıştır. Bu çalışmada da alanyazında bulunan diğer çalışmalar gibi protein fonksiyonlarının tahmini gerçekleştirilmiştir. Ancak, önerilen çalışma alanyazında belirtilen diğer çalışmalardan bazı yönleriyle farklıdır. Bu farklılıklar şu şekilde ifade edilebilir;

- Çalışmada sadece insana ait proteinler ve bunların fonksiyonları değerlendirilmiştir. Önceki çalışmaların çoğunda genellikle mayaya ve insana ait proteinler ve fonksiyonlarının üzerinde durulduğu görülmektedir,
- İncelenen protein fonksiyonlarının tahmini çalışmalarında protein anotasyonları genellikle n-gram, ProtVec, tek-vektör kodlama, k-mer gibi yöntemlerle sayısallaştırılmaktadır. Bu çalışmada bu yöntemlere başvurulamamıştır. Bunların aksine evrim-tabanlı (BLOSUM62, PAM250), fizikokimyasal-tabanlı (Atchley faktörleri), sinyal-tabanlı (EIIP) ve kaotik-tabanlı (FIBHASH) yöntemler kullanılmıştır,
- Önceki çalışmalar incelendiğinde, protein fonksiyonlarını tahmin edebilmek için protein-protein etkileşimlerinden ve protein bölgesi bilgilerinden de yararlanıldığı görülmüştür. Bu çalışmada böyle bir yapıya başvurulmamış ve sadece protein dizilimleri ile onların GO terimleri kullanılmıştır.

3 Materyal ve yöntemler

Bu kısımda çalışmada kullanılan protein anotasyonları ve GO terimleri hakkında bilgiler verilmiştir. Ayrıca, çalışma kapsamında kullanılmış olan protein haritalama teknikleri irdelenmiştir. Çalışmada önerilen yöntemin akış diyagramı Şekil 2'de verilmiştir.



Şekil 2. Çalışmada önerilen yöntemin akış diyagramı.

Figure 2. Flow chart of the method proposed in the study.

3.1 Protein verilerinin elde edilmesi

Çalışmada kullanılan protein anotasyonları UniProt veri kümesinden elde edilmiş ve FASTA formatı şeklinde değerlendirilmiştir [19]. Bunun yanı sıra çalışmada kullanılan GO terimleri ise GOA veri kümesinden elde edilmiştir [20]. GO terimleri 2013 ve öncesine ait terimlerdir. Tüm bu terimler deneysel olarak etiketlenmiş ve doğrulanmıştır. GO terimlerine dayalı bir protein fonksiyon tahmin işlemi gerçekleştirildiği için 3 kategoriye ait GO terimleri kullanılmıştır. BP kategorisi için toplamda 50 adet GO terimi belirlenmiş ve her bir terim için 40 anotasyon kullanılmıştır. MF için 50 adet GO terimi belirlenmiş ve bu kategori için 20 adet anotasyon kullanılmıştır. Son olarak CC kategorisi için 50 adet GO terimi ve 20 adet anotasyon kullanılmıştır. Toplamda 3 kategori için 4.000 adet protein anotasyonu kullanılırken, 150 adet GO terimi çalışmada değerlendirilmiştir. Protein GO terimlerinin belirlenmesinin ve protein anotasyonlarının toplanmasının ardından, protein dizilimleri çeşitli protein haritalama teknikleri ile sayısallaştırılmıştır.

3.2 Protein anotasyonlarının sayısallaştırılması

Bu çalışmada, protein anotasyonları çeşitli protein haritalama teknikleri ile sayısallaştırılmış ve protein fonksiyonlarının tahmini gerçekleştirilmiştir. Çalışma kapsamınca, kullanılan protein haritalama teknikleri Atchley faktörleri, BLOSUM62, EIIP, FIBHASH ve PAM250'dir. Bu kısımda, bahsedilen protein haritalama teknikleri hakkında bilgiler verilmiştir.

3.2.1 Atchley faktörleri

Bu protein haritalama yöntemi, protein dizilimlerinde karşılaşılan metrik problemini çözmek için önerilmiştir [21]. Protein varyasyonlarının büyük ve yorumlanabilir bileşenlerinden türetilmiş olup amino asit dizilimlerinin fizyokimyasal özelliklerini yansıtmaktadır. Toplamda polarite, ikincil yapı, moleküler hacim, kodon yoğunluğu ve proteinin elektrostatik yükü olmak üzere 5 adet faktör bulunmaktadır. Çalışmada 1 No.lu faktöre ait değerler kullanılmıştır. Tablo 1'de Atchley faktörleri verilmiştir.

Tablo 1. Amino asitlerin Atchley faktör değerleri (Faktör 1)

Table 1. Atchley factor values of amino acids (Factor 1).

Amino Asit Kodu	Sayısal İfadesi	Amino asit Kodu	Sayısal İfadesi
M	-0.663	Q	0.931
W	-0.595	S	-0.228
F	-1.006	A	-0.591
Y	0.260	N	0.945
P	0.189	G	-0.384
C	-1.343	R	1.538
T	-0.032	I	-1.239
H	0.336	D	1.050
V	-1.337	E	1.357
L	-1.019	K	1.831

Tablo 1'deki ifadeler göz önünde bulundurulduğunda, $S(n) = [M\ F\ V\ L\ R\ G]$ şeklinde bir protein dizilimi Atchley faktörü ile $S(n) = [-0.663 - 1.006 - 1.337 - 1.019\ 1.538 - 0.384]$ olacak şekilde haritalanır.

3.2.2 BLOSUM62 matrisi

Bu yöntem protein dizilimleri arasındaki benzerlikleri belirlemek için protein bloklarından türetilmiştir [22]. Genellikle evrimsel olarak farklı protein dizilimleri arasındaki hizalamalara skor vermek için kullanılmaktadır. BLOSUM45, BLOSUM80 gibi farklı tür sürümleri mevcuttur. Tüm BLOSUM matrisleri oluşturulurken gözlemlenen hizalamalardan yararlanılmıştır. Tablo 2'de BLOSUM62 matrisine ait sayısal değerler verilmiştir. Çalışmada BLOSUM62 matrisinin kullanılmasının en büyük nedeni, bu matrisin yıllarca standart bir matris olarak kabul edilmesidir [23].

Tablo 2. Amino asitlerin BLOSUM62 değerleri

Table 2. BLOSUM62 values of amino acids

Amino asit Kodu	Sayısal İfadesi	Amino Asit Kodu	Sayısal İfadesi
M	5	Q	5
W	11	S	4
F	6	A	4
Y	7	N	6
P	7	G	6
C	9	R	5
T	5	I	4
H	8	D	6
V	4	E	5
L	4	K	5

Tablo 2'deki ifadeler göz önünde bulundurulduğunda, $S(n) = [M\ F\ V\ L\ R\ G]$ şeklinde bir protein dizilimi BLOSUM62 matrisi ile $S(n) = [5\ 6\ 4\ 4\ 5\ 6]$ olacak şekilde haritalanır.

3.2.3 EIIP

EIIP protein haritalama yöntemi protein-protein ve protein-DNA arasındaki etkileşimleri belirlemek için önerilmiş bir yöntemdir [24]. Bu yöntem ile genomik dizilimler Fourier dönüşümü ile sinyallere ayrıştırılmış ve ayrıştırılmış olan sinyallerden güç spektrum değerleri elde edilmiştir. Ardından elde edilmiş olan güç spektrum değerleri her bir amino aside atanmış ve sayısallaştırma işlemi gerçekleştirilmiştir. Tablo 3'te her bir amino asit kodunun EIIP yöntemi ile sayısallaştırılmış değeri verilmiştir.

Tablo 3. Amino asitlerin EIIP değerleri.

Table 3. EIIP values of amino acids.

Amino asit Kodu	Sayısal İfadesi	Amino asit Kodu	Sayısal İfadesi
M	0.0823	Q	0.0761
W	0.0548	S	0.0829
F	0.0946	A	0.0373
Y	0.0516	N	0.0036
P	0.0198	G	0.0050
C	0.0829	R	0.0959
T	0.0941	I	0
H	0.0242	D	0.1263
V	0.0057	E	0.0058
L	0	K	0.0371

Tablo 3'teki ifadeler göz önünde bulundurulduğunda, $S(n) = [M\ F\ V\ L\ R\ G]$ şeklinde bir protein dizilimi EIIP yöntemi ile $S(n) = [0.0823\ 0.0946\ 0.0057\ 0\ 0.0959\ 0.0050]$ olacak şekilde haritalanır.

3.2.4 FIBHASH

FIBHASH protein haritalama yönteminde yazarlar protein dizilimlerini Fibonacci serisini ve hash tablosunu kullanarak sayısallaştırmışlardır [25]. Çalışmada önerilen yöntem protein ailelerini sınıflandırılmasında kullanılmış ve diğer yöntemler kadar etkili sonuçlar vermiştir. Proteinler alfabetik olarak sıralanmış ve Fibonacci dizisinin her bir elemanı bu amino asitlere atanmıştır. Doğada 20 çeşit amino asit olduğu için Fibonacci dizisinin 20. elemanına kadar seri açılmıştır.

Değerlerin yüksek olması nedeni ile araştırmacılar, değerleri hash tablosuna aktarmış ve değerlerin boyutlarını azaltarak protein dizilimlerini sayısallaştırmışlardır.

Tablo 4'te her bir amino asit kodunun FIBHASH yöntemi ile sayısallaştırılmış değeri verilmiştir.

Tablo 4. Amino asitlerin FIBHASH değerleri

Table 4. FIBHASH values of amino acids

Amino asit Kodu	Sayısal İfadesi	Amino asit Kodu	Sayısal İfadesi
M	9	Q	17
W	12	S	10
F	5	A	1
Y	19	N	7
P	16	G	8
C	2	R	0
T	18	I	6
H	13	D	3
V	11	E	4
L	15	K	14

Tablo 4'teki ifadeler göz önünde bulundurulduğunda, $S(n) = [M\ F\ V\ L\ R\ G]$ şeklinde bir protein dizilimi FIBHASH yöntemi ile $S(n) = [9\ 5\ 11\ 15\ 0\ 8]$ olacak şekilde haritalanır.

3.2.5 PAM250 matrisi

PAM250 matrisi dizilimler arasındaki benzerliği belirlemek için hizalanmış protein dizilimlerine puan vermek amacıyla geliştirilmiş bir modeldir [26]. Bu yöntem ile her bir amino asit koduna değerler nokta mutasyonlara göre verilmiştir. Tablo 5'te her bir amino asit kodunun PAM250 matrisi ile sayısallaştırılmış değeri verilmiştir.

Tablo 5. Amino asitlerin PAM250 değerleri.

Table 5. PAM250 values of amino acids.

Amino asit kodu	Sayısal ifadesi	Amino asit kodu	Sayısal ifadesi
M	0	Q	4
W	17	S	2
F	9	A	2
Y	10	N	2
P	6	G	5
C	12	R	6
T	3	I	5
H	6	D	4
V	4	E	4
L	6	K	5

Tablo 5'deki ifadeler göz önünde bulundurulduğunda, $S(n) = [M\ F\ V\ L\ R\ G]$ şeklinde bir protein dizilimi PAM250 matrisi ile $S(n) = [0\ 9\ 4\ 6\ 6\ 5]$ olacak şekilde haritalanır.

Protein dizilimleri belirtilen protein haritalama yöntemleri ile sayısallaştırıldıktan sonra BiLSTM ile protein fonksiyonları sınıflandırılmıştır.

4 Uygulama sonuçları

Bu kısımda çalışmada tasarlanmış olan BiLSTM hakkında bilgi verilmiş ve derin öğrenme algoritmasının parametreleri bu kısımda incelenmiştir. Bunlara ek olarak, her bir GO kategorisinin sınıflandırma başarım ölçütleri hesaplanmış ve her bir protein haritalama tekniğinin sonuçları incelenerek, kıyaslama yapılmıştır.

4.1 Protein GO terimlerinin sınıflandırılması

Günümüzde teknolojinin hızlı bir şekilde gelişmesiyle hemen hemen her alanda bilgisayar tabanlı sistemlere ihtiyaç duyulmaya başlanmıştır. İnsan gibi düşünebilen, insan gibi davranışlar sergileyebilen sistemlerin kullanımı yaygınlaşmıştır [44]. Çeşitli alanlarda kullanılan bu sistemler yapay öğrenme yöntemleri kullanılarak geliştirilmektedir. Yapay öğrenme denince ilk olarak akıllara makine öğrenmesi gelmektedir. Ancak, günümüzde veri sayısının hızlı bir şekilde artış göstermesi makine öğrenmesi tabanlı algoritmaların daha az kullanılmasına neden olabilmektedir [45].

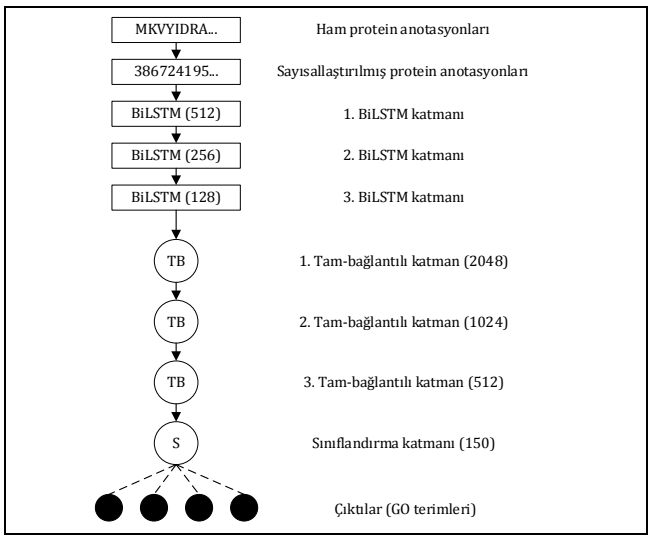
Makine öğrenmesi tabanlı uygulamalarda, genellikle örüntü tanıma teknikleri (özellik çıkarım) kullanılmaktadır. Veri sayısının çok olması durumunda bu süreç çok zaman almakta, insan gücü gerektirmekte ve elde edilen özellik vektörünün karmaşık olmasına neden olmaktadır [46].

Bunlara ek olarak ise, bazı çalışma alanlarında ise uzman bilgisi gerekmektedir. Bu gibi dezavantajlardan dolayı derin öğrenme popüler olmaya başlamıştır. Derin öğrenme ile özellik vektörü model tarafından elde edilmektedir. Bu da derin öğrenmenin makine öğrenmesine göre en büyük avantajıdır. Derin öğrenme günümüzde biyoenformatik uygulamalarında [47],[48], uzay teknolojilerinde [49], sağlık alanlarında [50],[51],[56], robot teknolojisinde [52], giyilebilir teknolojilerde [55], gerçek zamanlı sistemlerde [57] vb. alanlarda olmak üzere etkili bir şekilde kullanılmaktadır. Bu alanlarda elde edilen başarılardan dolayı, bu çalışmada derin öğrenme algoritması kullanılmıştır ve bunun için BiLSTM modeli tasarlanmıştır. RNN (Recurrent Neural Networks-Tekrarlayıcı Sinir Ağları) zaman serisi verilerini girdi olarak alıp sınıflandırma ve tahmin işlemi yapan bir çeşit derin öğrenme algoritmasıdır. Ayrıca ileri beslemeli bir algoritma olarak da değerlendirilir ve dâhili bir hafızaya sahiptir [27].

Genel olarak her tür girdi için aynı fonksiyonel işlemleri gerçekleştirirken, çıktı için geçmişteki hesaplamalar değerlendirilir. Hesaplama işleminin ardından çıktı tekrarlayıcı ağa geri gönderilir. Tekrarlayıcı sinir ağları ile ilgili detaylı bilgiler [28,29] No.lu çalışmalardan elde edilebilir. LSTM (Long-Short Term Memory - Uzun-Kısa Vadeli Bellek) sinir ağları, bir çeşit tekrarlayıcı sinir ağıdır. RNN mimarilerinde ortaya çıkan kaybolan gradyan problemine çözüm getirebilmek için önerilmiştir. Tıpkı RNN yapılarında olduğu gibi, zaman serilerinin sınıflandırılmasında ve tahmin işlemlerinde etkili olmaktadır [30],[31].BiLSTM ağları, LSTM ağlarının uzantısı olmakla beraber, dizi sınıflandırma problemlerini daha etkili bir şekilde çözebilmek için geliştirilmiştir [32],[34]. BiLSTM mimarisinde ileri ve geri yönlü hesaplamalar aynı anda değerlendirilir ve iki yönde yapılan işlemler sonucu elde edilen bilgiler birleştirilir ve çıktı elde edilir. Bu yapı sayesinde, iki yönde de bulunan bilgilerin sıralı verileri işlemede avantaj sağladığı gözlemlenmiştir [35]. Ayrıca, iki yönlü ağların tek yönlü ağlardan daha etkili olduğu belirlenmiştir [36]. İki yönlü mimarilerin daha etkili olmasından dolayı, bu çalışmada iki yönlü uzun-kısa vadeli bellek kullanılmıştır. Çalışmada çok katmanlı BiLSTM yapısı tasarlanmış ve sınıflandırıcının başarımı her bir protein haritalama tekniği için doğruluk, kesinlik, hassasiyet, en yüksek f skor ve AUC skorları ile ölçülmüştür. Tasarlanan BiLSTM yapısının parametreleri deneme yanılma yaklaşımı ile belirlenmiş ve en iyi sonucu veren parametrelerle tahmin işlemi gerçekleştirilmiştir. Bu parametreler şu şekilde ifade edilebilir;

- Girdi katmanında 4.000 adet protein anotasyonu kullanılmıştır. Sayısallaştırılmış olan bu protein anotasyonları başka herhangi bir önişleme tabi tutulmadan ikinci katmana gönderilmiştir,
- İkinci katmanda 512 üniteli BiLSTM kullanılmıştır. Aktivasyon fonksiyonu olarak ise SELU (Scaled Exponential Linear Units-Ölçekli Üstel Doğrusal Birimler) hesaplaması yapılmıştır,
- Ardından üçüncü katman tasarlanmış ve bu katmanda 256 üniteli BiLSTM kullanılmıştır. İlk uzun-kısa vadeli bellek katmanında olduğu gibi bu katmanda da aktivasyon fonksiyonu olarak SELU kullanılmıştır,
- Dördüncü katmanda 128 üniteli BiLSTM oluşturulmuştur. Bu katmanda da aktivasyon fonksiyonu olarak SELU hesaplamasından yararlanılmıştır,
- Ardından seyreltme işlemi yapılmış ve verilerin %25'inin unutulması sağlanmıştır. Bu işlemdeki ana amaç ağın ezberleme yapmasını engellemektir,
- Seyreltme işleminin ardından yığın normalleştirme işlemi yapılmıştır. Bu sayede verilerin hepsinin 0 ve 1 aralığında olması sağlanmıştır,
- Son olarak ise düzleştirme işlemi yapılmış ve matris formatındaki veri düzleştirilmiştir. Bu işlemin ana amacı verileri tam-bağlantılı katmana hazır hale getirmektir,
- Geliştirilen derin öğrenme modelinde üç adet tam bağlantılı katman oluşturulmuştur. Birinci tam bağlantılı katmanda 2048 adet nöron, ikinci tam-bağlantılı katmanda 1024 adet nöron ve son tam-bağlantılı katmanda ise 512 adet nöron kullanılmıştır,
- Son katmanda ise sınıflandırma yapılmıştır. Toplamda 150 adet GO terimi olduğu için bu katmanda 150 adet nöron kullanılmış ve aktivasyon fonksiyonu olarak ise Softmax'a başvurulmuştur,
- Modelin kaybı için ise kategorik çapraz entropi kullanılmıştır,
- En iyilime işlemi için ise RMSProp (Root Mean Square Propagation-Kök Ortalama Kare Yayılımı) kullanılmıştır,
- Döngü sayısı ise 500 olarak belirlenmiştir,
- Modeli doğrulamak için 5 katlı çapraz doğrulama yapılmıştır,
- Modeli test etmek için ise kör bir veri kümesi oluşturulmuştur. Test veri kümesinde BP, CC ve MF kategorileri için 30 adet GO terimi olmak üzere 90 adet sınıf belirlenmiştir. Her bir GO terimi için 10 adet olmak üzere toplamda 900 adet protein anotasyonu kullanılmıştır.

Şekil 3'te geliştirilen BiLSTM yapısı verilmiştir.



Şekil 3. Tasarlanan BiLSTM modeli.

Figure 3. Designed BiLSTM model.

4.2 Protein haritalama tekniklerinin başarım ölçütlerinin belirlenmesi

Protein haritalama tekniklerinin performansları doğruluk, kesinlik, hassasiyet, en yüksek f skor ve AUC skorları ile belirlenmiştir. Doğruluk skoru, eğitilen bir modelde doğru tahmin edilen değerlerin, veri kümesindeki toplam veri sayısına oranı ile belirlenmektedir. Doğruluk (D) skoru Denklem 1'de verilen formüle göre hesaplanır;

$$D = \frac{GP + GN}{GP + YP + GN + YN} \quad (1)$$

Denklem 1'de GP, gerçek pozitif değerini, GN gerçek negatif değerini, YP yalancı pozitif değerini ve YN ise yalancı negatif değerini ifade etmektedir.

Biyomedikal, sağlık, biyoinformatik gibi kritik alanlarda doğruluk skoru tek başına yeterli olmamaktadır. Bu tarz sistemlerde hassasiyet, kesinlik ve f-skor ve AUC skorları da önem arz etmektedir [53,54]. Kesinlik (K) başarım ölçütü pozitif olarak tahmin edilen değerlerin gerçekten kaç tanesinin pozitif olduğunu göstermektedir. Denklem 2'de verilen formüle göre hesaplanır.

$$K = \frac{GP}{GP + YP} \quad (2)$$

Bunun yanı sıra, hassasiyet (H) ise pozitif olarak tahmin edilmesi gereken değerlerin kaç tanesini pozitif olarak tahmin edildiğini göstermektedir. Denklem 3'te verilen formül ile hesaplanır.

$$H = \frac{GP}{GP + YN} \quad (3)$$

En yüksek f-skor (F) ise kesinlik ve hassasiyet değerlerinin harmonik ortalaması ile elde edilmektedir. Denklem 4'te verilen formül ile hesaplanır.

$$F = 2 * \frac{K * H}{K + H} \quad (4)$$

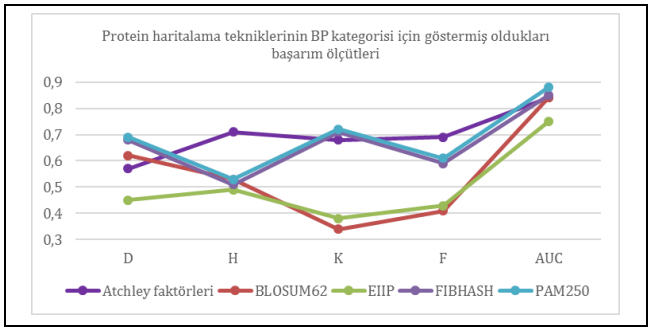
AUC skoru ise eğitilmiş olan modelin sınıfları ne kadar başarılı bir şekilde ayırt edebildiğini göstermek için kullanılmaktadır. AUC skoru arttıkça, model doğru girdileri doğru sınıflara aktaracaktır. AUC skoru eğri altında kalan kısmın alanın hesaplanmasıyla elde edilir ve hesaplama işlemi için Simpson'un Kuralı uygulanmaktadır. Tablo 6'da tüm GO kategorileri için elde edilmiş ve her bir protein haritalama tekniği için hesaplanmış olan başarım ölçütleri verilmiştir. Şekil 4, 5 ve 6'da bu yöntemlerin her bir GO terimi için göstermiş olduğu sonuçlar görsel olarak verilmiştir. Tablo 6'daki sonuçlar incelendiğinde, her bir GO kategorisi için farklı sonuçlar elde edildiği açıkça görülmektedir. BP kategorisinde, en etkisiz sonucu EIIP protein haritalama tekniği göstermiş ve doğruluk %45 olarak hesaplanmıştır. Bu sonucu Atchley faktörleri yöntemi izlemiş ve %57 oranında bir doğruluk gözlemlenmiştir. BLOSUM62, FIBHASH ve PAM250 protein haritalama yöntemleri, EIIP ve Atchley faktörlerine nazaran daha başarılı bir doğruluk sonucu üretmişler ve her bir haritalama tekniği %60'ın üzerinde performans sergilemiştir. BP kategorisi için en iyi performans PAM250 matrisi ile elde edilmiştir. Doğruluk sonuçlarının yanı sıra, tüm yöntemlerin AUC skorlarında başarılı bir performans söz konusudur. AUC skorunun 1 olması demek, yöntemin protein fonksiyonlarını belirlemede etkili olduğunu ifade etmektedir.

Tablo 6. Protein haritalama tekniklerinin her bir GO kategorisi için göstermiş oldukları başarım ölçütleri.

Table 6. Performance criteria of protein mapping techniques for each GO category.

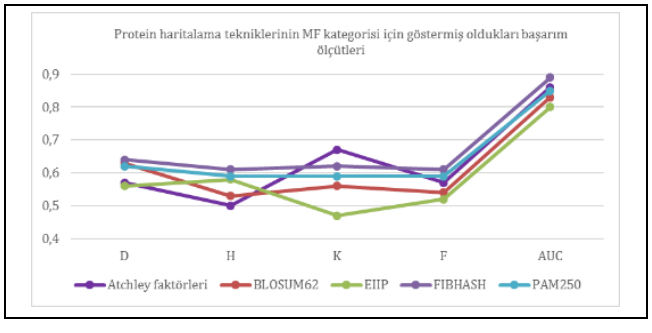
Protein Haritalama Teknikleri	BP					MF					CC				
	D	H	K	F	AUC	D	H	K	F	AUC	D	H	K	F	AUC
Atchley Faktörleri	0.57	**0.71**	0.68	**0.69**	0.84	0.57	0.50	**0.67**	0.57	0.86	**0.64**	**0.71**	0.64	**0.67**	0.87
BLOSUM62	0.62	0.53	0.34	0.41	0.84	0.63	0.53	0.56	0.54	0.83	0.62	0.66	0.64	0.65	0.85
EIIP	0.45	0.49	0.38	0.43	0.75	0.56	0.58	0.47	0.52	0.80	0.52	0.61	0.43	0.50	0.80
FIBHASH	0.68	0.51	0.71	0.59	0.85	**0.64**	**0.61**	0.62	**0.61**	**0.89**	**0.64**	0.63	**0.67**	0.65	**0.88**
PAM250	**0.69**	0.53	**0.72**	0.61	**0.88**	0.62	0.59	0.59	0.59	0.85	0.60	0.51	0.51	0.51	0.85

Biyoenformatik ve medikal çalışmalarda AUC skoru büyük bir önem arz etmektedir [37]-[39]. Bir sınıflandırıcının AUC skoru %70'den büyükse o sınıflandırıcı makul bir sınıflandırıcı olarak kabul edilmektedir [40]. %80 ve %90 arasındaki değerler harika sınıflandırıcı olarak ifade edilebilirken, değerin %90'dan büyük olması sınıflandırıcının muhteşem performans gösterdiğini belirtmektedir [41]. Bu bilgiler ışığında, EIIP yönteminin makul bir sınıflandırma yaptığı, geri kalan yöntemlerin ise harika bir sınıflandırma yaptığı gözlemlenmiştir.



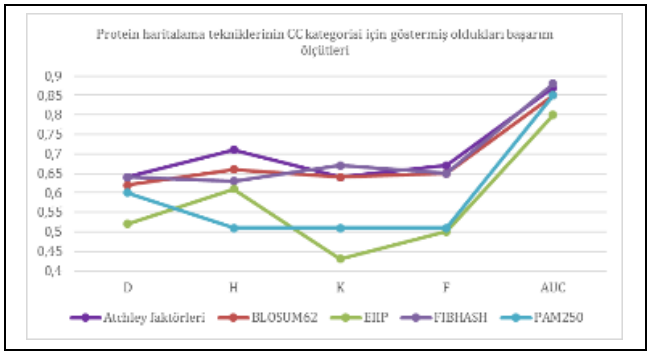
Şekil 4. Protein haritalama tekniklerinin BP GO terimi için göstermiş oldukları başarım ölçütleri (X ekseni başarım ölçütlerini, Y ekseni ise başarım ölçütlerinin değerlerini ifade etmektedir)

Figure 4. Performance criteria of protein mapping techniques for the BP GO term (While the X axis expresses the performance evaluation criteria, the Y axis refers to the values of the performance evaluation criteria.



Şekil 5. Protein haritalama tekniklerinin MF GO terimi için göstermiş oldukları başarım ölçütleri (X ekseni başarım ölçütlerini, Y ekseni ise başarım ölçütlerinin değerlerini ifade etmektedir).

Figure 5. Performance criteria of protein mapping techniques for the MF GO term (While the X axis expresses the performance evaluation criteria, the Y axis refers to the values of the performance evaluation criteria.



Şekil 6. Protein haritalama tekniklerinin CC GO terimi için göstermiş oldukları başarım ölçütleri (X ekseni başarım ölçütlerini, Y ekseni ise başarım ölçütlerinin değerlerini ifade etmektedir).

Figure 6. Performance criteria of protein mapping techniques for the CC GO term (While the X axis expresses the performance evaluation criteria, the Y axis refers to the values of the performance evaluation criteria.

En yüksek AUC skoru, doğruluk skorun da olduğu gibi PAM250 matrisi ile elde edilmiştir. MF kategorisi içinde benzer bir çıkarım yapılabilir. Bu GO kategorisinde EIIP yöntemi BP kategorisinde olduğu gibi en etkisiz yöntem olmuş ve %56 oranında bir doğruluk sonucu üretmiştir. BP kategorisindeki başarımının aksine, Atchley faktörleri MF için etkisiz bir doğruluk sonucuna ulaşmıştır. BLOSUM62, FIBHASH ve PAM250 protein haritalama yöntemleri ise en etkili yöntemler olmuştur. Bu kategoride en yüksek doğruluk sonucu %64 ile FIBHASH yöntemi ile elde edilmiştir. Bunun yanı sıra AUC skorlarında ise gözle görülür bir başarı söz konusudur. Bu kategoride tüm protein haritalama yöntemleri başarılı bir sınıflandırma işlemi gerçekleştirmiştir. Bu kategoride ise en yüksek AUC skoru %89 ile FIBHASH yöntemi ile elde edilmiştir. CC kategorisinde ise, diğer GO kategorilerinde olduğu gibi en etkisiz yöntem EIIP yöntemi olmuştur. Diğer tüm protein haritalama yöntemleri %60'ın üzerinde bir doğruluk skoruna erişmiştir. Bu kategoride en yüksek doğruluk sonucu Atchley faktörleri ve FIBHASH yöntemi ile elde edilmiş ve sonuçlar %64 olarak hesaplanmıştır. Benzer şekilde en yüksek AUC skorları tekrar bu yöntemler ile elde edilmiş ve sırasıyla %87 ve %88 oranında değerler gözlemlenmiştir. Tıpkı MF kategorisinde olduğu gibi, bu kategoride de tüm protein haritalama teknikleri %80 ve %90 arasında AUC skoru ürettiği için, harika sınıflandırma performansı sergilemiştir. Sonuçlarda da anlaşıldığı üzere, çalışmada kullanılan protein haritalama tekniklerinin genel olarak sınıflandırma işleminde etkili olduğu gözlemlenmiştir.

5 Tartışma

Üç farklı GO kategorisinde EIIP yönteminin en etkisiz olduğu gözlemlenmiştir. Bu durumun en büyük nedeni EIIP protein haritalama tekniğinde ortaya çıkan bozulma (dejenerasyon) olabilir. Bozulma işlemi farklı amino asitlerin aynı sayısal değerleri almasıyla ortaya çıkmaktadır [42]. Atchley faktörleri ve FIBHASH yöntemlerinde böyle bir durum söz konusu değildir. PAM250 ve BLOSUM62 matrislerinde de, EIIP yönteminde olduğu gibi farklı amino asitlerin benzer değer aldıkları bilinmektedir. Ancak, bu yöntemlerin EIIP yönteminden farkı protein dizilimlerinin belirli bir protein bilgisine dayanarak sayısallaştırılmasıdır. PAM250 ve BLOSUM62 matrisleri evrim-tabanlı yöntemlerdir. Bu yöntemlerde protein anotasyonları, proteinlerin evrim bilgisine dayanarak sayısallaştırılmıştır [43]. EIIP protein sayısallaştırma yönteminde böyle bir durum söz konusu değildir. Proteinler bu yöntemde sinyal işleme ile sayısallaştırılmaktadır. Bu yaklaşım ile bilgiler kaybolmuş olabilir. EIIP dışındaki diğer yöntemler hemen hemen benzer sonuçlar üretmişlerdir. Doğruluk ve AUC skorları birbirine yakındır. Çalışmanın sonunda sonuçlar birbirine yakın da olsa, protein haritalama tekniklerinin protein fonksiyonlarını belirlemede önemli bir rol oynadığı görülmüştür. Tasarlanmış olan derin öğrenme modeli ile AUC skorlarına göre tüm yöntemlerin her GO kategorisi için başarılı bir sınıflandırma yaptığı belirlenmiştir. Çalışmanın avantajları şu şekilde ifade edilebilir;

- Bu çalışma ile protein haritalama tekniklerinin protein fonksiyonlarını belirlemede başarılı olduğu gözlemlenmiştir. Bundan dolayı hesaplama-tabanlı yaklaşımların bu alanda kullanılabileceği gözlemlenmiştir. Alanyazında bulunan mevcut çalışmalar bu sonucu desteklemektedir,
- Protein haritalama tekniklerine ek olarak, derin öğrenme algoritmalarının çoğu biyoenformatik çalışmalarında da olduğu gibi, bu alanda da etkili ve başarılı olduğu gözlemlenmiştir. Makine öğrenmesinde ortaya çıkan hatalar, derin öğrenme yöntemleri ile giderilmiş ve bu çalışma ile bu çıkarım doğrulanmıştır.

Çalışmadaki avantajların yanı sıra dezavantajlar da mevcuttur. Bu dezavantajlar şu şekilde sıralanabilir;

- Alanyazında bulunan diğer çalışmalara nazaran, bu çalışmada kullanılan veriler daha azdır. Bu durumun nedeni donanım ihtiyacımızın yetersiz olmasından kaynaklanmaktadır. Veri sayısının arttırılmasıyla bu sonuçlarda iyileşme olabilir ya da daha kötü performans sonuçları elde edilebilir. Bu durumun ileriki çalışmalarda değerlendirilmesi gerekmektedir,
- Farklı bir derin öğrenme algoritması ile sonuçlar farklı bir şekilde değerlendirilebilir. Çalışmanın çeşitliliğini ve başarımını koruması için, diğer derin öğrenme algoritmaları ile de değerlendirilmesi gerekmektedir,
- Bu çalışmada kullanılan protein haritalama teknikleri dışında farklı protein haritalama tekniklerinin kullanılması çalışmanın doğruluğu ve güvenilirliği açısından önem arz etmektedir. Diğer protein haritalama teknikeri ile de çalışmanın değerlendirilmesi bu bakımdan önemlidir,
- Alanyazındaki çalışmalar incelendiğinde, protein fonksiyonlarının doğruluk sonuçlarının düşük olduğu gözlemlenmektedir. Bu çalışma da sonuçlar incelendiğinde doğruluk, kesinlik, hassasiyet ve en yüksek f skorunun düşük olduğu görülmüştür. Bu durumun en büyük nedenlerinden birisi çok sayıda GO terimi olmasıdır. GO teriminin çok olması sınıflandırıcının etiket ayrımını doğruluk gibi değerlendirme metrikleri açısından başarılı bir şekilde yapmasına izin vermemektedir.

6 Sonuçlar

Bu çalışmada GO terimlerine dayalı protein fonksiyonu tahmini gerçekleştirilmiştir. Çalışma kapsamında farklı protein haritalama teknikleri kullanılmış ve bu protein haritalama tekniklerinin performansları doğruluk, hassasiyet, kesinlik, en yüksek f skor ve AUC skorları ile belirlenmiştir. Çalışmada BP, MF ve CC olmak üzere üç farklı GO kategorisi kullanılmıştır. BP kategorisi için 40 adet protein anotasyonu, MF ve CC kategorisi için ise 20 adet protein anotasyonu olmak üzere toplamda 80 adet protein anotasyonu kullanılmıştır. Çalışmada değerlendirilen protein anotasyonları UniProt veri kümesinden elde edilmiştir. Sınıflandırma işlemi BiLSTM derin öğrenme modeli tasarlanmış ve tasarlanan model üzerinde GO terimleri tahmin edilmiştir. Sınıflandırıcının doğrulama işlemi için 5 katlı çapraz-doğrulama kullanılmıştır. Doğrulama işleminin ardından kör bir veri kümesi hazırlanmış ve sınıflandırıcının ve protein haritalama tekniklerinin başarımı bu veri kümesi üzerinden test edilmiştir. Test veri kümesi için her bir GO kategorisinde 300 protein anotasyonu olmak üzere 900 adet protein anotasyonu değerlendirilmiştir. Bunlara ek olarak test verisinde toplamda 90 adet GO terimi bulunmaktadır. Çalışmanın sonunda her bir GO kategorisi için protein haritalama tekniklerinin başarımları karşılaştırılmıştır. BP kategorisi için en başarılı sınıflandırma PAM250 matrisi ile elde edilmiştir. Bu yöntem ile %69 oranında doğruluk ve %88 oranında AUC skoru elde edilmiştir. MF kategorisinde ise en başarılı sonuç FIBHASH ile elde edilmiştir. FIBHASH ile %64 doğruluk ve %89 AUC skorları elde edilmiştir. MF kategorisinde olduğu gibi CC kategorisinde de en başarılı yöntem FIBHASH yöntemi olmuş ve sırasıyla doğruluk ve AUC için % 64 ve %89 oranında değerler gözlemlenmiştir. Çalışmanın sonunda tasarlanan derin öğrenme modelinin ve seçilen protein haritalama yönteminin, protein fonksiyonlarını belirlemede etkili rol oynadıkları belirlenmiştir.

7 Conclusions

In this study, protein function prediction based on GO terms has been performed. Within the scope of the study, different protein mapping techniques were used and the performances of these protein mapping techniques were determined with accuracy, recall, precision, highest f score and AUC scores. Three different GO categories were used in the study including, BP, MF and CC. A total of 80 protein annotations were used, including 40 protein annotations for the BP category and 20 protein annotations for the MF and CC categories. The protein annotations used in the study were obtained from the UniProt data set. For the classification process, a bidirectional long-short-term memory deep learning model was designed and GO terms were predicted on the designed model. 5-fold cross-validation was used for the validation of the classifier. After the validation process, a blind data set was prepared and the performance of the classifier and protein mapping techniques

was tested on this data set. For the test data set, 900 protein annotations, including 300 protein annotations in each GO category, were evaluated. In addition to these, there were a total of 90 GO terms in the test data. At the end of the study, the performance of protein mapping techniques for each GO category was compared. The most successful classification for the BP category was obtained with the PAM250 matrix. With this method, 69% accuracy and 88% AUC score were obtained. In the MF category, the most successful result was obtained with FIBHASH. 64% accuracy and 89% AUC scores were obtained with FIBHASH. As in the MF category, the most successful method in the CC category was the FIBHASH method, and values of 64% and 89% were observed for accuracy and AUC, respectively. At the end of the study, it was determined that the designed deep learning model and the selected protein mapping method play an effective role in determining protein functions.

8 Kaynaklar

[1] Rifaigolu AS, Dogan T, Martin MJ, Cetin-Atalay R, Atalay V. "DEEPred: Automated protein function prediction with multi-task feed-forward deep neural networks". *Scientific Reports,* 9(1), 7344, 2019.

[2] Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC. "The genomes on line database (GOLD) in 2009: Status if genomic and metagenomica projects and their associated metadata". *Nucleic Acids Research,* 38, 346-354, 2010.

[3] Cao R, Cheng J. "Integrated protein function prediction by mining function associations, sequences, and protein-protein and gene-gene interaction networks". *Methods,* 93, 84-91, 2016.

[4] The UniProt Consortium. "UniProt: the universal protein knowledgebase". *Nucleic Acids Research,* 45, 158-169, 2017.

[5] Bonetta R, Valentino G. "Machine learning techniques for protein function prediction". *Proteins,* 88(3), 397-413, 2020.

[6] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. "Gene ontology: tool for the unification of biology". *Nature Genetics,* 25(1), 25-29, 2000.

[7] Friedberg I. "Automated protein function prediction: the genomic challenge." *Briefings in Bioinformatics,* 7(3), 225-242, 2006.

[8] Lee D, Redfern O, Orengo C. "Predicting protein function from sequence and structure". *Nature Reviews: Molecular Cell Biology,* 8(12), 995-1005, 2007.

[9] Bernardes JS, Pedreira CE. "A review of protein function prediction under machine learning perspective". *Recent Patents on Biotechnology,* 7(2), 122-141, 2013.

[10] Fa R, Cozzetto D, Wan C, Jones DT. "Predicting human protein function with multi-task deep neural networks". *PLOS One,* 13(6), 1-6, 2018.

[11] Lobley AE, Nugent T, Orengo CA, Jones DT. "FFPred: an integrated feature-based function prediction server for vertebrate proteomes". *Nucleic Acids Research,* 36, 297-302, 2008.

[12] Suthaharan S. "Big data classification: problems and challenges in network intrusion prediction with machine learning". *ACM SIGMETRICS Performance Evaluation Review,* 41(4), 70-73, 2014.

[13] Najafabadi MM,Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. "Deep learning applications and challenges in big data analytics". *Journal of Big Data,* 2(1), 1-21, 2015.

[14] Cai Y, Wang J, Deng L. "SDN2GO: an integrated deep learning model for protein function prediction". *Frontiers in Bioengineering,* 8, 1-11, 2020.

[15] Cao R, Freitas C, Chan L, Sun M, Jiang H, Chen Z. "ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network". *Molecules,* 22(10), 1-14, 2017.

[16] f M, Khan MA, Hoehndorf R, Wren J. "DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier". *Bioinformatics,* 34(4), 660-668, 2018.

[17] You R, Zhang Z, Xiong Y, Sun F, Mamitsuka H, Zhu S. "GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank". *Bioinformatics,* 34(14), 2465-2473, 2018.

[18] Hakala K, Kaewphan S, Bjorne J, Mehryary F, Moen H, Tolvanen M, Salakoski T, Ginter F. "Neural network and random forest medels in protein function prediction". *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 18, 1-10, 2020.

[19] UniProt Consortium. "UniProt: a hub for protein information". *Nucleic Acids Research,* 43, 204-212, 2015.

[20] Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apwiler R. "The GOA database in 2009: an integrated gene ontology annotation resource". *Nucleic Acids Research,* 37, 396-403, 2009.

[21] Atchley WR, Zhao J, Fernandes AD, Drüke T. "Solving the protein sequence metric problem". *Proceedings of the National Academy of Sciences of the United States of America,* 102(18), 6395-6400, 2005.

[22] Henikoff S, Henikoff JG. "Amino acid substitution matrices from protein blocks". *Proceedings of the National Academy of Sciences of the United States of America,* 89(22), 10915-10919, 1992.

[23] Styczynski MP, Jensen KL, Rigoutsos I, Stephanopoulos G. "BLOSUM62 miscalculations improve search performance". *Nature Biotechnology,* 26(3), 274-275, 2008.

[24] Veljkovic N, Glisic S, Prljic J, Perovic V, Botta M, Veljkovic V. "Discovery of new therapeutic targets by the informational spectrum method". *Current Protein & Peptide Science,* 9(5), 493-506, 2008.

[25] Alakus TB, Turkoglu I. "A novel fibonacci hash method for protein family identification by using recurrent neural networks". *Turkish Journal of Electrical Engineering & Computer Sciences,* 29(1), 370-386, 2021.

[26] Dayhoff MO, Schwartz RM, Orcutt BC. "A model of evolutionary change in proteins". *National Biomedical Research Foundation,* 5(3), 345-352, 1978.

[27] Can B. "LSTM ağları ile Türkçe kök bulma". *Bilişim Teknolojileri Dergisi,* 12(3), 183-193, 2019.

[28] Şeker A, Diri B, Balık H. "Derin öğrenme yöntemlerin ve uygulamaları hakkında bir inceleme". *Gazi Mühendislik Bilimleri Dergisi,* 3(3), 47-64, 2017.

[29] Metin İA, Karasulu B. "İnsan aktivitelerinin sınıflandırılmasında tekrarlayan sinir ağı kullanan derin öğrenme tabanlı yaklaşım". *Veri Bilimi Dergisi,* 2(2), 1-10, 2019.

[30] Sherstinsky A. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network". *PhysicaD: Nonlinear Phenomena,* 404, 1-28, 2020.
[31] Hochreiter S, Schmidhuber J. "Long short-term memory". *Neural Computation,* 9(8), 1735-1780, 1997.
[32] Liu G, Guo J. "Bidirectional LSTM with attention mechanism and convolutional layer for text classification". *Neurocomputing*, 337, 325-338, 2019.
[33] Basaldella M, Antolli E, Serra G, Tasso C. "Bidirectional LSTM recurrent neural network for keyphrase extraction". *Italian Research Conference on Digital Libraries,* Udine, Italy, 25-26 January 2018.
[34] Graves A, Jaitly N, Mohamed A. "Hybrid speech recognition with deep bidirectional LSTM". *2013 IEEE Workshop on Automatic Speech Recognition and Understanding,* Olomouc, Czech Republic, 8-12 December 2013.
[35] Babüroğlu B, Tekerek A, Tekerek M. "Türkçe için derin öğrenme tabanlı doğal dil işleme modeli geliştirilmesi". *arXiv,* 2019. https://arxiv.org/pdf/1905.05699.pdf
[36] Graves A, Schmidhuber J. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". *Neural Networks,* 18(5-6), 602-610, 2005.
[37] Kamarudin AN, Cox T, Kolamunnage-Dona R. "Time-dependent ROC curve analysis in medical research: current methods and applications". *BMC Medical Research Methodology,* 17(53), 1-19, 2017.
[38] Safari S, Baratloo A, Elfil M, Negida A. "Evidance based emergency medicine; part 5 receiver operating curve and area under the curve". *Emergency,* 4(2), 111-113, 2016.
[39] Zhao XG, Dai W, Li Y, Tian L. "AUC-based biomarker ensemble with an application on gene scores predicting low bone mineral density". *Bioinformatics,* 27(21), 3050-3055, 2011.
[40] Wigton RS, Connor JL, Centor RM. "Transportability of a decision rule for the diagnosis of streptococcal pharyngitis". *Archives of Internal Medicine,* 146(1), 81-83, 1986.
[41] Mandrekar JN. "Receiver operating characteristic curve in diagnostic test assessment". *Journal of Thoracic Oncology,* 5(9), 1315-1316, 2010.
[42] Chen D, Wang J, Yan M, Bao FS. "A complex prime numerical representation of amino acids for protein function comparison". *Journal of Computational Biology,* 23(8), 669-677, 2016.
[43] Jing X, Dong Q, Hong D, Lu R. "Amino acid encoding methods for protein sequences: a comprehensive review and assessment". *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 17(6), 1918-1931, 2020.
[44] Doğan F, Türkoğlu İ. "Derin öğrenme modelleri ve uygulama alanlarına ilişkin bir derleme". *DÜMF Mühendislik Dergisi,* 10(2), 409-445, 2019.
[45] Alpaydın E. *Yapay Öğrenme.* 4. Baskı. İstanbul, Türkiye, Boğaziçi Üniversitesi, 2018.
[46] Goodfellow I, Bengio Y, Courville A. *Derin Öğrenme.* 1. Baskı. Ankara, Türkiye, Buzdağı, 2018.
[47] Das B, Turkoglu I. "A novel numerical mapping method based on Entropy for digitizing DNA sequences". *Neural Computings and Applications,* 29(8), 207-215, 2018.
[48] Alakus TB, Turkoglu I. ". A novel Entropy-based mapping method for determining the protein-protein interactions in viral genomes by using coevolution analysis". *Biomedical Signal Processing and Control,* 65, 1-15, 2021.
[49] Dogan F, Turkoglu I. "Classification of satellite images by deep learning". *8th International Advanced Technologies Symposium,* Elazig, Turkey, 19-22 October 2017.
[50] Alakus TB, Turkoglu I. "Comparison of deep learning approaches to predict COVID-19 infection". *Chaos, Solutions & Fractals,* 140, 1-7, 2020.
[51] Toraman S, Alakus TB, Turkoglu I. "Convolutional capsnet: A novel artificial neural network approach to detect COVID-19 disease from X-ray images using capsule networks". *Chaos, Solutions & Fractals,* 140, 1-11, 2020.
[52] Gurgoze G, Turkoglu I. "Energy management techniques in mobile robots". *World Academy of Science, Engineering and Technology International Journal of Energy and Power Engineering,* 10(11), 1079-1087, 2017.
[53] Pala MA, Çimen ME, Boyraz ÖF. "Meme kanseri teşhis edilmesinde karar ağacı ve knn algoritmalarının karşılaştırmalı başarım analizi". *Academic Perspective Procedia,* 2(3), 544-552, 2019.
[54] Kim J, Kim J, Lee D, Chung KY. "Ontology driven interactive healthcare with wearable sensors". *Multimedia Tools and Applications,* 71, 827-847, 2014.
[55] Iskanderov J, Güvensan MA. "Akıllı telefon ve giyilebilir cihazlarla aktivite tanıma: klasik yaklaşımlar, yeni çözümler". *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi,* 25(2), 223-239, 2019.
[56] Gürkan H, Hanilçi A. "Evrişimsel sinir ağı ve QRS imgeleri kullanarak EKG tabanlı biyometrik tanıma yöntemi". *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi,* 26(2), 318-327, 2020.
[57] Çetin M, Beyhan S, Bahtiyar B. "Yapay sinir ağı temelli uyarlamalı doğrusal model-öngörülü kontrol". *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi,* 22(8), 650-658, 2016.
[58] Vascon S, Frasca M, Tripodi R, Valentini G, Pelillo M. "Protein function prediction as a graph-transduction game". *Pattern Recognition Letters,* 134, 96-105, 2020.
[59] Makrodimitris S, van Ham RCHJ, Reinders MJT. "Improving protein function prediction using protein sequence and GO-term similarities". *Bioinformatics,* 35(7), 1116-1124, 2019.
[60] Gligorijevic V, Barot M, Bonneau R. "deepNF: deep network fusion for protein function prediction". *Bioinformatics,* 34(22), 3873-3881, 2018.