Yapay Sinir Ağları Kullanılarak Protein Katlanması Tanıma

Araştırma Makalesi/Research Article

Sena DİKİCİ*, Volkan ALTUNTAŞ

Bilgisayar Mühendisliği, Bursa Teknik Üniversitesi, Bursa, Türkiye

sena.dikici@btu.edu.tr, valtuntas@gmail.com

(Geliş/Received:06.07.2022; Kabul/Accepted:06.03.2023)

DOI: 10.17671/gazibtd.1141468

Özet— Proteinler uzun aminoasit zincirlerinden oluşur ve vücut kimyasını düzenlemekle birlikte hücrelerin yapısı ve aralarındaki iletişim için öneme sahiptir. Bir proteinin hücre bazındaki görevini gerçekleştirebilmesi için, molekülü hücredeki hedefiyle etkileşime girebilecek üç boyutlu yapıya dönüştüren bir bükülme süreci olan katlanma işlemini gerçekleştirmesi gerekir. Sıcaklık, ağır metaller veya kimyasal durumlar gibi etkenler proteinlerin yanlış katlanmasına sebep olabilir. Yanlış katlanan proteinler, vücuttaki görevini yerine getiremez. Alzaymır, kistik fibrozis, deli dana hastalığı gibi hastalıklara sebep olabilir. Protein katlanması tanıma işlemi, biyologlar açısından bir problem olarak değerlendirilir. Literatürde yer alan şablon tabanlı yaklaşımlara karşın yapay sinir ağları, protein katlanması probleminin çözümüne yönelik yüksek başarım gösterir. Yapay sinir ağları, ele alınan problemin çözümü için geniş veri kümelerinde yer alan ve problemin çözümüne katkı sağlayacak bilgi kazancı yüksek özellikleri kullanan bir hesaplama tekniğidir. Bu çalışmada SCOPe 2.06, SCOPe 2.07, SCOPe 2.08 veri setleri kullanılarak şablon tabanlı yaklaşımlardan elde edilen sonuçların yapay sinir ağı yöntemi ile birleştirilerek protein katlanması tanıma işlemi gerçekleştirilmiştir. Gerçekleştirilen deneyler sonucunda yapay sinir ağı yönteminin katkısı ile literatürde yer alan sonuçların iyileştirildiği görülmüştür. Bu çalışma ile biyoinformatik alanında protein katlanması tanıma probleminin çözümüne yeni bir yaklaşım sunularak literatüre katkı sağlanması amaçlanmıştır.

Anahtar Kelimeler— protein katlanması, yapay sinir ağları, protein katlanması tanıma, biyoinformatik

Protein Folding Recognition by Artificial Neural Networks

Abstract— Proteins are made up of amino acid chains and are important for the structure of the cells and their communication with each other, while regulating body chemistry. For a protein to perform its function on a cell basis, it must perform the folding process that converts the molecule into a three-dimensional structure that can interact with its target in the cell. Factors such as temperature, heavy metals or chemical conditions may cause proteins to be folded incorrectly. Incorrectly folded proteins cannot perform their role in the body. Protein folding recognition is considered a problem for biologists. Despite the template-based approaches in the literature, artificial neural networks show high performance in solving the protein folding problem. Artificial neural networks are a computational technique that uses high-specification knowledge in large data sets to help solve the problem that is being addressed and will contribute to the solution of the problem. In this study, protein folding recognition was performed by combining the results from template-based approaches with artificial neural network method using the Scope 2.06, Scope 2.07, Scope 2.08 datasets. The results in the literature were improved by the contribution of the artificial neural network method because of the experiments conducted. This study aims to contribute literature by introducing an innovative approach to the solution of the problem of protein folding recognition in the field of bioinformatics.

Keywords—protein folding, artificial neural network, protein folding recognition, bioinformatics

1. GİRİŞ (INTRODUCTION)

Proteinler, organizmaların faaliyetlerinde önemli rol oynayan bileşenlerdir. Bir proteinin işlevi, diğer proteinlerle etkileşime girmesi ve katlanmasıyla belirlenir. Proteinlerin görevleri ve birbirleriyle etkileşimleri, proteinlerin yapılarının sınıflandırılması açısından önemlidir. Proteinlerin üçüncül yapılarının tanımlanması, protein fonksiyon gösterimi, ilaç tasarımı, moleküler biyoloji gibi alanlar için en önemli görevlerden biridir. RNA (Ribonükleit asit) 'ya kopyalanan genetik bilginin bir protein veya protein zinciri haline dönüşmesi sırasınad, her protein lineer bir amino asit zinciri veya sabit bir üç boyutlu yapısı olmayan rastgele bir bobin olarak sentezlenir. Zincirdeki amino asitler, sonunda iyi tanımlanmış, katlanmış bir protein oluşturmak için birbirleriyle etkileşime girer. Aminoasit dizileri proteinlerin üç boyutlu yapılarını belirler. Protein katlanması, proteinlerin dizileri ile yapıları arasındaki ilişkiyi açıklar. Bir polipeptit zincirinin biyolojik olarak aktif bir protein haline gelmesi için katlandığı sürece protein katlanması denir. Proteinlerin katlanma yapıları, fonksiyonlarının tanınması için önemli rol oynar. Amino asit dizilerindeki geniş çeşitlilik ise protein yapısındaki farklı konformasyonları açıklar. Mevcut durumdaki konformasyonda proteinler kararlılığını sürdürdüğü takdirde kendilerine ait görevleri yerine getirebilmektedir. Bu görevler biyolojik işlemlerin doğru şekilde devamlılığını sağlar.

Bir proteinin genel katlanma türünün karakterize edilmesi önemli bir etkendir. Proteinin yapısal sınıfının ilk tanımı 1976 yılında Levitt ve Chothia tarafından yapılmıştır. Birinci aşama (α), az miktarda iplikten oluşur. İkinci aşamada (β), proteinler dört yapısal sınıftan ve az sayıda sarmaldan oluşur. Üçüncü aşama (α/β), sarmallardan ve paralel dizilerden oluşur. Son aşama (α+β) ise sarmallardan ve paralel olmayan dizilerden oluşur [1]. Proteinlerin yapısal sınıfı veya katlanma türü, moleküler biyoloji alanında önem taşır. Proteinlerin, dizi kimliklerinden bağımsız olarak katlanma türü sayesinde üç boyutlu yapısı tespit edilebilmektedir [2].

Sıcaklık, ağır metaller veya kimyasal olaylar gibi etkenler proteinlerin yanlış katlanmasına sebep olmaktadır. Proteinlerin yanlış katlanması, hastalıklara sebep olabilmektedir. Örneğin, bazı kanser türleri p53 ismi verilen proteinin, bu protein kanser proteini olarak da adlandırılmaktadır, yanlış katlanması sebebiyle meydana gelmektedir [3]. Beyin hücrelerindeki proteinlerin yanlış katlanması ve kontrol edilemez boyuta ulaşması sonucunda alzaymır ve deli dana hastalığı meydana gelmektedir. Protein katlanmasının doğru tahmini, yanlış katlanma sonucu meydana gelen hastalıkların erken ve hızlı teşhis edilmesi açısından önem taşımaktadır.

Protein katlanması tanıma, hedef proteinlerin sadece sekans bilgilerine dayanarak bilinen protein yapı şablonlarına göre sınıflandırmayı amaçlayan bir problemdir. Şablon tabanlı yaklaşımlardan elde edilen sonuçların iyileştirilmesi amacıyla bu yaklaşımlar makine öğrenmesi ya da derin öğrenme yöntemleri ile birleştirilebilmektedir. Çalışmamızda da protein katlanması türlerinin, yapay sinir ağları kullanılarak tespit edilmesi ile literatürde yer alan sonuçların iyileştirilmesi amaçlanmıştır. Çalışmamızın protein katlanması tanıma için zaman ve doğruluk değerleri kriterleri açısından olumlu yönde katkı sağlayacağı öngörülmektedir. Hedef ve sorgu protein dizileri arasında şablon tabanlı yaklaşımlardan elde edilen üç farklı benzerlik puanı yöntemine ait sonuçlar kullanılmıştır. Benzerlik puanları, yapay sinir ağları ile eğitilerek elde edilen sonuçlar literatüre sunulmuştur.

2. LİTERATÜR (RELATED WORK)

Biyoloji ve tıp alanında yapılan çalışmaların günümüze yaklaştıkça bilgisayar bilimleri yöntemleri ile birleştirildiği çalışmaların yaygınlaştığı görülmüştür [4,5,6,7]. Protein katlanması tanıma problemi için bilgisayar bilimleri ile birleştirilerek yapılan çok sayıda çalışma mevcuttur. Protein katlanması tanıma işlemi proteinlerin 3 boyutlu yapılarıyla ilişkilendirilmektedir. Bu doğrultuda protein yapısının protein sekansları üzerinden tahmin edilmesi, problemin önemi ve iyi tanımlanmış hesaplamalı temelleri sebebiyle uzun süredir araştırmacılar tarafından incelenmektedir. Bu çalışmalar bilgisayar bilimleri ile birleştirilerek yeni bir çalışma alanı oluşturmuştur. Makine öğrenmesi yöntemleri ile protein yapısının tahmin edilmesi ile ilgili çalışmalar yapılmıştır. Derin artık ağlar protein yapılarının tahmini için yaygın olarak kullanılan yöntemlerden biridir. Örneğin, ResNet mimarisi birçok çalışmada yer alan sistemler için taban kabul edilen mimari olmuştur. Çalışmalar sonucunda ResNet mimarisinin güçlü olduğu ve sonuçları iyileştirdiği literatüre sunulmuştur [8]. Protein katlanması tanıma problemi için matematiksel temellere dayalı geleneksel yöntemler literatürde yer almaktadır. Belirli algoritmalar kullanılarak şablon tabanlı yaklaşımlar ile katlanma türü tanıma işlemi gerçekleştirilmiştir. Bu yaklaşıma sahip çalışmalardan birinde yerel uyarlanabilir komşu bağlantısı (local adaptive neighbor connection) yöntemi kullanılmıştır. 23 farklı protein kullanılan çalışmada yerel uyarlanabilir komşu bağlantısı yönteminin 18 protein için en yüksek kalite değerine sahip olduğu, 12 protein için ise ya en hızlı ya da en hızlı ikinci performansa sahip olduğu görülmüştür [9]. Makine öğrenmesi yöntemleriyle gerçekleştirilen çalışmaların başarısı sayesinde biyoinformatik alanı literatürde ve araştırmalarda güncelliğini korumaktadır. Protein katlanması tanıma problemi de bu kapsamda yer almaktadır. Protein katlanması ile ilgili çalışmalardan birinde SPARKS-X (Diziden Yapı Tahmini - Structure Prediction from Sequence), HHblits (HMM-HMM tabanlı ışık hızında yinelemeli dizi arama - HMM-HMM-based lightning-fast iterative sequence search) ve DeepFR (Derin Katlanma Tanıma - Deep Fold Recognition) şablon tabanlı yaklaşımlarından benzerlik puanları elde edilmiştir. SPARKS-X, sorgunun tahmin edilen tek boyutlu yapısal özellikleri ile şablonların karşılık gelen doğal özellikleri arasında olasılığa dayalı eşleştirmeyi kullanarak gelişen bir protein katlama tanıma ve şablon tabanlı modellemedir. HHblits ise protein yapısının ve işlevinin diziye dayalı

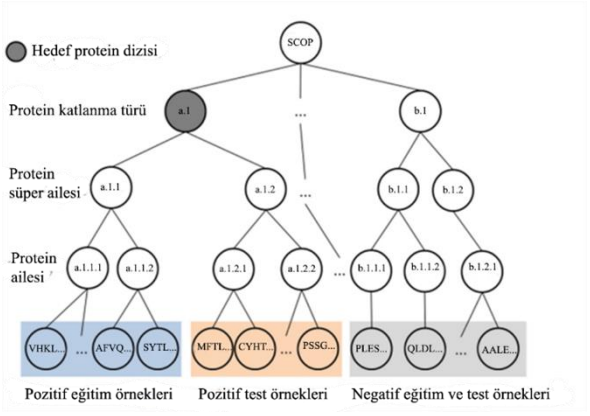
doğru tahmini için geliştirilmiş Gizli Markov modeli tabanlı yinelemeli bir dizi arama yöntemidir. DeepFR ise süper aile ve katlanma seviyelerinde protein katlanması tanıma için derin evrişimli sinir ağı kullanan bir yaklaşımdır [10]. Elde edilen benzerlik puanları, global bir özellik matrisinde birleştirilmiştir. Proteinlerin dizi yapılarından elde edilen benzerlik matrisi, destek vektör makinesine eğitim verisi olarak verilmiştir. Bu yaklaşıma TSVM-fold ismi verilmiştir. Destek vektör makinesi yöntemi için farklı çekirdek fonksiyonları kullanılarak yapılan çalışmada en yüksek doğruluk değeri %63 olarak elde edilmiştir [11]. Şablon tabanlı yaklaşımlarla protein katlanması tanıma konulu başka bir çalışmada ise SPARKS, FOLDpro, SPARKS-X ve BoostThreader yöntemleri kullanılmıştır. Çalışmada kullanılan şablon tabanlı yaklaşımlardan elde edilen doğruluk değerleri sırasıyla %47.7, %48.8, %67, %57.4 olarak elde edilmiştir [12]. Makigaki ve Ishida'nın gerçekleştirmiş olduğu çalışmada FFAS (Katlama ve İşlev Atama Sistemi - The Fold and Function Assignment System), HHPRED ve SPARKS-X şablon tabanlı yaklaşımlar kullanılarak sırasıyla %47, %47, %52 benzerlik puanı elde edilmiştir [13]. HHPRED, protein homoloji tespiti ve yapı tahmini için etkileşimli sunucudur [14]. Bin Liu ve diğerlerinin çalışmasında ise çoklu dizi hizalaması yaklaşımı, destek vektör makinesi yöntemi ile birleştirilmiştir. SCOP veri tabanının 1.67 versiyonu kullanılmış ve %78 doğruluk değeri elde edilmiştir [15] Sudha ve diğerleri EDD, DD, TG, RDD veri setleri üzerinde yapay sinir ağı yöntemini kullanarak protein yapısal sınıflandırması ve katlanması tanıma için yeni bir yöntem önermişlerdir. Daha önce yapılan çalışmalardan elde edilen özniteliklerin farklı kombinasyonları kullanılarak yapay sinir ağı yöntemini, destek vektör makinesi ve Bayes yöntemi ile elde ettikleri deneysel sonuçlarla karşılaştırarak yapay sinir ağı yönteminin yüksek başarım elde ettiği sonucunu sunmuşlardır. EDD, DD, TG ve RDD veri setlerine ait veriler yapay sinir ağları ile eğitildiğinde elde edilen doğruluk değerleri sırasıyla %83, %76.6, %76, %73.3 olmuştur [2]. Wei ve Zou'nun çalışmasında n adet özellik temsili için n adet tekli sınıflandırıcı eğitilerek topluluk sınıflandırıcı stratejileriyle birleştirilmiştir. Çalışmada 5 farklı veri seti topluluk sınıflandırıcı yöntemi ile protein katlanma sınıflarına ayrılarak elde edilen sonuçlar karşılaştırılmıştır. Çalışma sonucunda DD veri setinin dengesiz olduğu, SCOP veri tabanının ise fazla sayıda katlanma türü içermesi sebebiyle kısıtlar oluşturduğu sunulmuştur. ProFold sistemi, topluluk sınıflandırıcı yöntemi kullanılarak %76.2 değer ile en yüksek doğruluk değerine sahip olmuştur [16]. Başka bir çalışmada ise proteinlere ait aminoasit özelliklerinin farklı kombinasyonları kullanılarak çalışmanın başarımı değerlendirilmiştir. Bu çalışmada göreli temas sırası modeli ve sabit parametreler kullanılarak, mutlak temas siparişi modeli ve sabit parametreler kullanılarak, bağıl temas sırası modeli ve değişken parametreler kullanılarak, mutlak temas sırası modeli ve değişken parametreler kullanılarak farklı deneyler gerçekleştirilmiştir. Bu deneyler sonucunda en yüksek korelasyon değerine sahip deney mutlak teması sırası modeli ve değişken parametreler kullanılarak elde edilmiştir [17]. Protein katlanması ve dinamiği ile ilgili çalışmalardan bazıları verilerden elde edilen bilgilerin makine öğrenmesi yöntemleri kullanılarak simüle edilebilmesine yöneliktir. Proteinlerdeki yarı kararlı durumların ve bunların istatistiklerinin çıkarılması için Markov Durum Modelleri (Markov State Models) kullanılmaktadır. Protein katlanması ve açılmasının hızlı sümilasyonu, protein dinamikleri hakkında da hızlı bir şekilde bilgi edinilmesini sağlamaktadır. Yapılan çalışma makine öğrenmesi yöntemlerinin sıklıkla kullanılmasa da çalışma sonuçlarında iyileştirme sağladığını göstermiştir [18].

3. MATERYAL VE METODLAR (MATERIAL AND METHODS)

Bu bölümde çalışmada kullanılan veri seti, veri setlerinde mevcut olan ve çalışmada kullanılan öznitelikler, çalışmada kullanılan yapay sinir ağları yöntemi açıklanmıştır. Yapay sinir ağları yöntemine ait fonksiyonlar ve algoritmalar açıklanmıştır. Modelin başarım ölçütlerine yer verilmiştir.

3.1. Veri Seti (Dataset)

Çalışmamızda SCOPe veri tabanında yer alan veri setleri kullanılmıştır. SCOPe, Berkeley Laboratuvarlarında geliştirilen bir veri tabanıdır. Proteinlerin yapısal sınıflandırılması amacıyla oluşturulmuş olup proteinlerin yapısal ilişkilerine ait veriler yer almaktadır. SCOPe veri tabanı, protein yapılarını ve dizilerini analiz etmek için oluşturulmuştur. Proteinler arasındaki yapısal ve evrimsel ilişkilerin doğru, ayrıntılı ve kapsamlı bir tanımını sağlamayı amaçlamaktadır. Protein akrabalıkları ve protein kıvrımlarına ilişkin geniş bir araştırma sonucu oluşturulan SCOPe veri tabanı, özellikle protein dizilerinden tanınmayacak kadar eski olan veriler hakkında güvenilir bilgiler sağlayarak, araştırmalar ve sınıflandırma için bir çerçeve oluşturmaktadır [19]. Bu çalışmada SCOPe veri tabanının 2.06, 2.07 ve 2.08 versiyonu kullanılmıştır. SCOPe 2.06 veri setinde 244.332 adet veri ve 1431 katlanma türü, SCOPe 2.07 veri setinde 276.231 adet veri ve 1457 adet katlanma türü, SCOPe 2.08 veri setinde 344.848 adet veri ve 1485 adet katlanma türü mevcuttur.



Şekil 1. SCOP veri setinin sınıf yapısı [11].
(The structure of SCOP dataset)

Şekil 1'de SCOPe veri setinin sınıf yapısına göre protein katlanma türü, protein süper ailesi ve protein ailesi sınıflandırmasının nasıl yapıldığına ilişkin gösterim yer almaktadır.

3.2. Özellik Seçimi (Feature Selection)

Özellik seçimi mevcut veri kümesinin en iyi temsilini oluşturan alt kümenin belirlenmesi olarak tanımlanmaktadır. Veri tabanındaki özellik sayısı azaltılarak veri boyutunun azaltılmasını sağlar. Özellik seçimi için bilgi kazanımı (information gain) yöntemi kullanılmıştır. Karar ağaçları, kök düğümden başlanarak aşağıya doğru karar verici yönlendirmeler ile oluşturulan ağaçlardır. Ağacı inşa etmek için entropi kullanılmaktadır. Aslında kullanılan yöntemi entropi oluşturmaktadır. Entropiyi kullanan bu yönteme bilgi kazanımı denmektedir. Entropi, tanımlandığı her alandaki düzensizliğin ölçülmesi için kullanılmaktadır. Denklem 1'de entropi formülü verilmiştir. Entropi H sembolüyle gösterilmektedir. H, bir bilginin bir ortamdaki düzensizliğini belirlemek için kullanılmaktadır. Bu noktada bilginin aynı tek düze yapıda akması düzensizliği ortadan kaldırır. Entropi formülünde n verinin miktarını ve p_i ise ilgili verinin olma olasılığını ifade etmektedir. Bilgi kazanımı ise durumlar arası entropi farkından oluşmaktadır. Denklem 2'de ise bilgi kazanımına ait formül verilmiştir.

$$H = -\sum_{i=1}^{n} p_i \log(p_i) \tag{1}$$

$$\begin{aligned} \text{Bilgi Kazanımı } (S,A) &= H(S) - \sum_{t \in T} p(t)H(t) \\ &= H(S) - H(S \mid A) \end{aligned} \tag{2}$$

3.2.1. DM Puanı (DM Score)

Çalışmada eğitilmek üzere, SCOPe veri setlerinde yer alan proteinlere ait özelliklerden bilgi kazanımı sonucunda yüksek puana sahip özellikler seçilmiştir. Bu özelliklerden biri DM puanıdır.

Protein diziliminin analizi ve yapılarının karşılaştırılması ile protein katlanma türü sınıflandırılır [20]. Katlanma türü, proteinlerin ikinci aşamada karşılık gelen yapılarının kararlığına göre sınıflandırılmıştır. Korunan ortak çekirdeğin çoklu yapısal hizalaması ile karşılık gelen dizilerin korelasyonu tartışılmıştır.

Bir proteinin yapısı, proteinlere ait moleküllerin farklı kısımlarında katlanma sonucu göstereceği değişikliğe ait atomik koordinatların listesi sayesinde elde edilmektedir [21]. Bir proteine ait yüzeyin bağıl konumu, proteinin göreviyle yakından ilişkilidir; yüzeye ait parçalar diğer moleküllerle doğrudan etkileşime girer. Proteinlerin yapısı itibarıyla birbirlerinden farklı görevleri gerçekleştiren grupların vermiş olduğu kimyasal tepkimeler, proteinlerin tepkimeye girdiği arayüzle ilişkilerine bağlıdır. Lee ve Richard algoritmasında protein yapısında bulunan her bir atom veya atom grubu, tanımlandığı boyutta (çözücü veya çözünen) bir moleküle erişebilme yeteneğinin listesi sayesinde düzensiz yüzeyin açıklaması yapılır [21]. Hesaplamalar baz alındığında titreşimler ya da esneklik hakkında hiçbir bilgi yansıtmayan proteinin koordinat listesinde, türetilen sayılar statik erişilebilirliğe sahiptir. Bu tür sayıların, kimyasal verilerin yorumlanması için koordinatların kendisinden daha yararlı olacağı öne sürülmüştür.

Geliştirilen bilgisayar programları ile bir protein molekülünün içinde en az bir çözücü molekülü barındıracak büyüklükteki boşluklar tespit edilebilmektedir. Bu, protein bölümlerinde içbükey eğriliğe sahip tüm erişilebilirlik yüzeylerini yalıtarak ve bu içbükey yüzeyin içinden molekülün dışına giden bir kanala sahip olanları ortadan kaldırarak yapılmaktadır [19]. Aminoasit kalıntılarının temsil ettiği kümenin bir alt kümesini belirlemek için, her bir koordinat veri kümesinde analiz gerçekleştirilmiştir. Kimyasal çözücünün erişebildiği temas yüzey alanı, Richmond ve Richards'ın (1978) algoritması kullanılarak, o katlama birimi çevresindeki her atom için hesaplanmıştır. Katlama ve işlev için gerekli olan ligandlar, merkezi bir metale bağlanan bir atom veya molekül, erişilebilirlik hesaplamasına dahil edilip gruplara bağlı inhibitörler dahil edilmemiştir [21].

Her atom için angstrom karesi cinsinden erişilebilirlik hesaplaması elde edildikten sonra, ana zincir ve yan zincir için kalıntılar toplanmıştır. Yan zincir boyutundaki varyasyon ve dolayısıyla yüzey alanındaki varyasyonlar, genişletilmiş bir konformasyon için erişilebilir yüzey alanı teorik olarak hesaplanmıştır ve toplamlar normalleştirilmiştir [21]. Daha sonra çözücü molekülle etkileşime giren yan zincirin yüzdelik değerleri elde edilmiştir [22].

Lee ve Richards algoritması ile Hubbard ve Blundell yaklaşımı birleştirilerek proteinlere ait DM puanı elde edilmiştir [22].

3.2.2. SP Puanı (SP Score)

Bilgi kazanımı denklemi (Denklem 2) ile hesaplama yapıldıktan sonra yüksek puana sahip olarak seçilen diğer özellik ise SP puanıdır. Bu yaklaşımda protein dizilerinin karşılaştırılması için uzak homolog proteinlerin, protein veri bankasında yer alan; moleküllerin yapısal olarak bilinen protein alanları ve zincirleri için hesaplanan yapısal hizalamalarının tutulduğu DASH veri tabanı kullanılmıştır. Ölçüm sistemleri analizi kullanılarak SP puanı hesaplanıp literatürde yer alan diğer ölçümlere göre %10 civarında iyileşme elde edildiği görülmüştür [23]. DASH hizalamaları önceden hesaplandığından ek hesaplama maliyeti ve ek ağ yükünden tasarruf sağlanmıştır. MAFFT-DASH yaklaşımı, son kullanıcı yükünü azaltarak sıralama ve yapısal hizalama bilgilerini entegre etmenin oldukça

uygun ve verimli bir yolunu sunmaktadır. Ayrıca düşük hesaplama maliyetleri ile doğru hizalamalar sağlamaktadır.

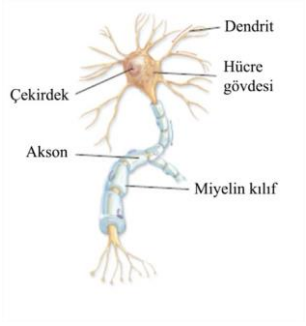
3.2.3. CL Puanı (CL Score)

Bilgi kazanımı yöntemi ile en yüksek puana sahip olan üçüncü özellik CL puanı olmuştur. CL puanı çalışmada kullanılmak üzere seçilmiştir. DM puanı, SP puanı ve CL puanları kullanılarak protein katlanması tanıma işlemi gerçekleştirilmiştir.

Biyokimya alanında bir biyomoleküle bağlanarak kompleks yapı oluşturan bileşiklere ligand denir. Ligand moleküller genellikle hedef proteinlerdeki bağlanma yerlerine iyonik bağlar, hidrojen bağları veya Van der Waals güçleri ile bağlanır [24]. Protein-ligand kompleksleri ise kristal yapılarındaki artış ile kanonik olmayan etkileşimlerin anlaşılmasına yardımcı olmuştur. Bu etkileşimler sayesinde kristal yapı veri tabanlarının kapsamlı analizi ile kanonik olmayan etkileşimlerinin yeni bir puanlama sistemiyle yeni bilgiler edinilmesi ve fonksiyonlarının geliştirilmesi sağlanmıştır. Yapılan çalışma sayesinde ligandlarda Cl grubunun aromatik halkaya bağlandığı bulunmuştur [25]. Protein-ligand komplekslerinin Cl etkileşimleri analiz edilerek protein veri bankası için yeni bir puanlama yöntemi ortaya konmuştur. Bu sayede etkileşimlerin doğası ve Cl etkileşimlerinin geometrisi anlaşılır hale gelmiştir [25]. Cl puanı sayesinde protein katlanması sırasında oluşan kompleks yapıların anlaşılır kılınması amaçlanmıştır.

3.3. Yapay Sinir Ağları (Artificial Neural Network)

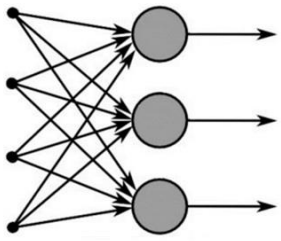
Yapay sinir ağları, insan beyninin çalışma şeklinin bilgisayar ortamına uyarlanması olarak tanımlanır. İnsanlara benzer şekilde deneyim kazanarak ve uygun metotlarla öğrenir. Yapay sinir ağları, verilerdeki kalıpları ve ilişkileri tespit ederek bilgileri toplar [26]. Nöronlar, hücre işleyişinin kontrolünü sağlayan çekirdek, hücreye bilgi taşıyan dentrit ve sinyali uzağa taşıyan akson olarak adlandırılan hücre gövdesinden oluşmaktadır (Şekil 2). İmpulslar akson boyunca sinapsa, bir nöron ile bir sonraki arasındaki bağlantıya geçer ve sinyaller birinden diğerine ya hep ya hiç şeklinde iletilir [26]. Nöronlar, birbirlerine tamamen bağlıdır ve öğrenme, tahmin etme ve tanıma yeteneğine sahiptir [26].



Şekil 2. Nöron yapısı [27].
(Structure of a neuron)

Yapay sinir ağı, sinir yapısını oluşturan katsayılar (ağırlıklar) ile bağlantılı yüzlerce tek birimden, yapay nöronlardan oluşan biyolojik olarak esinlenilmiş bir hesaplama modelidir. Yapay sinir ağında bulunan her bir verinin kendisine ait ağırlıkları, transfer edilme süreci ve ürettiği sonuç (çıktı) değeri vardır. Yapay sinir ağları girdi katmanında bulunan verilere bağlı olarak çıktı üreten bir denklem olarak da tanımlanır. Bağlantı ağırlıkları sistemin hafızasını temsil ettiği için yapay sinir ağlarına bağlantıcı modeller de denir. Tasarlanmış birçok sinir ağı türü vardır ve hepsi nöronlarının transfer fonksiyonları, öğrenme kuralı ve bağlantı formülü ile tanımlanabilir [26].

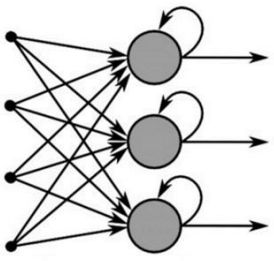
Yapay nöronlar uyarıcı veya engelleyici girdiler alabilir. Uyarıcı girdiler, bir sonraki nöronun toplama mekanizmasının eklenmesine neden olurken, engelleyici girdiler çıkarılmasına neden olur. Bir nöron aynı katmandaki diğer nöronları da inhibe edebilir. Buna lateral inhibisyon denir. Ağ, en yüksek olasılığı seçmek ve diğer olasılıkları engeller [28].



Şekil 3. İleri beslemeli ağ [29].
(Feedforward network)

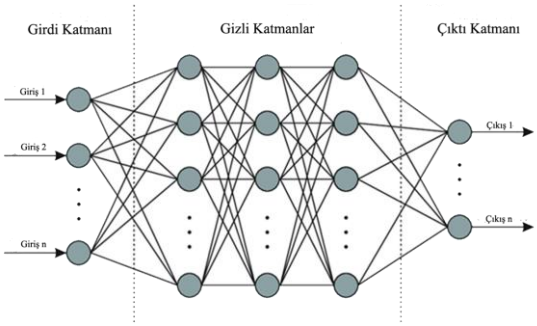
Geri bildirim, bir katmanın çıktısının bir önceki katmanın girişine veya aynı katmana geri yönlendirildiği başka bir bağlantı türüdür. İleri beslemeli mimaride ise girdi katmanından çıktı katmanına doğru ileri bağlantılar mevcuttur. Geri dönüş yapılmaz. Bu nedenle önceki çıktı değerlerinin kaydı tutulmaz (Şekil 3) [29].

Geri besleme mimarisi ise ileri besleme mimarisinin aksine çıktı katmanından girdi katmanında bulunan nöronlarla bağlantı kurabilir. Her nöronun eğitim sırasında belirli bir hata değerine sahip olması olasıdır. Geri beslemeli ağlar önceki durumda elde edilen sonuçları hafızasında tutar (Şekil 4). Önceki sonuçların hafızada tutulması sebebiyle bir sonraki durum mevcut giriş sinyallerinin yanı sıra ağın önceki durumlarına bağlıdır [29].



Şekil 4. Geri beslemeli ağ [29].
(Feedback network)

Çalışmamızda kullanılan yapay sinir ağı yapısı bir adet giriş katmanı, 3 adet gizli katman ve bir adet çıktı katmanından oluşmaktadır. Şekil 5'te temsili gösterimi verilmiştir. Nöron ve parametre sayıları Tablo 1'de verilmiştir.



Şekil 5. Yapılan çalışmanın yapay sinir ağı mimarisinin şematik gösterimi [30]
(Schematic representation of the artificial neural network architecture of the study)

3.3.1. Aktivasyon Fonksiyonu (Activation Function)

Tarihine bakıldığında, sigmoid fonksiyonu, çok katmanlı algılayıcı ve Boltzmann Makinesi gibi yapay sinir ağlarında nöronların aktivasyon işlevi olarak sıklıkla kullanılmıştır [30]. Sigmoid fonksiyonu veya tanjant hiperbolik fonksiyon gibi aktivasyon fonksiyonlarında hata değeri minimum hale getirilmek istendiğinde gradyan değeri sıfır olabilmektedir. Gradyan değerinin sıfır olması problemine, kaybolan gradyan problemi denmektedir. Kısıtlı Boltzmann Makinesini yöntemine ait bu problemin çözümüne yönelik olarak Nair ve arkadaşları doğrultulmuş doğrusal birimleri (ReLU) tanıttı [31].

Glorot, gizli katmanlar için uygulanan aktivasyon fonksiyonunun ReLU fonksiyonu olarak tercih edildiğinde, fonksiyonun kullanıldığı derin sinir ağlarının öğrenme hızının arttığını söylemiştir. ReLU, derin sinir ağları için standart olarak kullanılır [32]. Kaybolan gradyan sorununun ReLU aktivasyon fonksiyonu kullanılarak önlenebileceği görülmüştür [32].

ReLU fonksiyonunun temel avantajı aynı anda tüm nöronları aktive etmemesidir. Bir nöron, negatif değer ürettiğinde ReLU fonksiyonu nöronun aktive edilmemesini sağlamaktadır. Bu sebeple çok katmanlı sinir ağlarında ReLU aktivasyon fonksiyonu kullanılabilmektedir. Çalışmamızda geliştirilen model ise giriş katmanı, 3 adet gizli katman ve çıktı katmanından oluşması sebebiyle ReLU aktivasyon fonksiyonu kullanılmıştır.

3.3.2. Kayıp Fonksiyonu (Loss Function)

Yapay sinir ağlarının başarımını değerlendirme yöntemlerinden biri kayıp fonksiyonunun değeridir. Eğitilen modelin tahmin ettiği değerler ile gerçekte bilinen değerler arasındaki fark ile ifade edilir. Elde edilen değer sayesinde modelin ne kadar geliştirebileceği hakkında bilgi sahibi olunmasını sağlar [33].

Çalışmamızda kayıp fonksiyonu olarak kategorik çapraz düzensizlik fonksiyonu (categorical cross entropy) kullanılmıştır. Çapraz entropi hata fonksiyonu, eğitim esnasında oluşabilecek yavaşlamaların önüne geçmek amacıyla kullanılan bir fonksiyondur.

Aktivasyon fonksiyonlarının ardından kayıp fonksiyonları kullanılmaktadır. Çalışmamızda çoklu sınıflandırma gerçekleştirildiği için kategorik çapraz entropi (KÇE) kayıp fonksiyonu kullanılmıştır. Aktivasyon fonksiyonları çıktı olarak her bir sınıfın olasılık değerini verir. Fonksiyon çıktısı sıfır ile bir arasında değer almaktadır. Bu eşitlik kullanılarak sıfırdan büyük ve ağ çıktısının beklenen değere yakın olduğu durumlarda entropinin sıfıra yakın olduğu görülmektedir [34].

$$\text{KÇE} = -\sum_{q=1}^{l}\sum_{k=1}^{p} d_{qk}\log(y_{qk}) \qquad (3)$$

Denklem 3'te kategorik çapraz entropi kayıp fonksiyonu verilmiştir. Denklemde yer alan p, sınıf etiketi sayısını; l ise veri sayısını temsil etmektedir. Fonksiyon çıktıları 0 ile 1 aralığında olduğundan y_{qk}, [0,1] aralığındadır. d_{qk} değerleri gerçek değerleri, y_{qk} değerleri tahmin değerlerini temsil etmektedir.

3.3.3. *Optimizasyon Algoritması (Optimization Algorithm)*

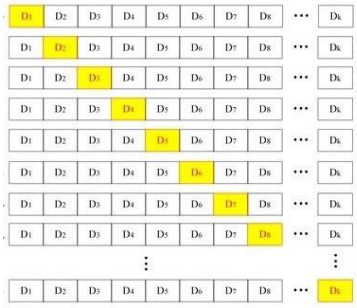
Yapay sinir ağlarının hata oranını en aza indirgemek için optimizasyon algoritmaları kullanılmaktadır. Öğrenme işlemi, filtre katsayıları ve katmanlar arası ağırlık değerlerinin veriyi temsil eden uzayında optimum olarak belirlenmesiyle gerçekleşmektedir. Çalışmamızda, optimum çözüme ulaşmada iyi performans gösteren Adam optimizasyon algoritması, öğrenme algoritması olarak seçilmiştir [35].

3.4. K-Katlamalı Çapraz Doğrulama (K-Fold Cross Validation)

K-Katlamalı Çapraz doğrulama yöntemi modelin değerlendirilmesinde kullanılan yöntemlerden biridir. Aşırı öğrenme (over fitting) ve eksik öğrenmenin (under fitting) tespit edilmesini sağlamaktadır. Aşırı öğrenme meydana geldiğinde model eğitim veri seti içerisindeki fonksiyonel örüntünün yanı sıra gürültülü verileri de öğrenir. Buna ezberleme de denmektedir. Ezberleme ise modeli eğitim veri setine bağlı kılar bu sebeple model yeni veriler üzerinde başarılı tahminde bulunamaz [36]. Eksik öğrenme ise modelin veri setindeki örüntüleri eksik öğrenmesinden kaynaklanmaktadır. Eksik öğrenme modelin genelleme yapmasına sebep olur. Aşırı öğrenme ve eksik öğrenme zayıflıklarının önüne geçmek için k-katlamalı çapraz doğrulama yöntemi kullanılmaktadır [37].

Şekil 6'daki akış şeması, k-katlı çapraz doğrulamanın, verileri rastgele K gruplara bölerek başladığını ve ardından her grup için aşağıdaki işlemlerin gerçekleştirildiğini göstermektedir [38].

Öncelikle veri seti üzerinde K adet parçaya bölünmüş veriler içerisinden bir adet alt küme belirlenerek test kümesi oluşturulur. Kalan K-1 parça ise eğitim kümesi olarak kullanılır [38]. Bu işlem K kez tekrarlanır. Her iterasyonda sıradaki alt küme test verisi olarak kullanılır.



Şekil 6. K-Katlamalı çapraz doğrulama süreci gösterimi (Sarı renkli kareler test veri setini, beyaz renkli kareler eğitim veri setini temsil etmektedir.) [38]
(The main process of K-fold cross-validation)

3.5. Başarım Ölçütü (Performance Measure)

Yapay sinir ağlarının başarımını değerlendirmek için kullanılan ölçütlerden biri de doğruluk değeridir. Doğru sınıflandırılan örnek sayısının toplam örnek sayısına bölünmesi ile doğruluk değeri elde edilir. Yanlış sınıflandırılmış örnek sayısının toplam örnek sayısında bölünmesi ile de hata oranı elde edilir.

$$\text{Doğruluk} = \frac{\text{DP+DN}}{\text{DP+YP+DN+YN}} \quad (4)$$

$$\text{Hata oranı} = \frac{\text{YP+YN}}{\text{DP+YP+DN+YN}} \quad (5)$$

Denklem 1 ve Denklem 2'de yer alan DP, doğru pozitif; DN, doğru negatif; YP, yanlış pozitif ve YN, yanlış negatif anlamlarını taşır.

4. BULGULAR (RESULTS)

Çalışmamızda SCOPe 2.06, SCOPe 2.07 ve SCOPe 2.08 veri setleri yapay sinir ağı yöntemleri kullanılarak eğitilmiştir. Değerlendirme ölçütü olarak K-katlamalı çapraz doğrulama yöntemine ait doğruluk değerleri alınmıştır. Katlama sayısı için 5, 10, 15 ve 20 değerleri ile deneyler gerçekleştirilmiştir. Gerçekleştirilen deneyler sonucunda doğruluk değeri en yüksek K değeri seçilmiştir.

Çalışmamızda kullanılan modelde yapay sinir ağlarına ait nöron sayısı, parametre sayısı, kayıp fonksiyonu, aktivasyon fonksiyonu ve optimizasyon algoritması Tablo 1'de verilmiştir.

Tablo 1. Scope 2.06 veri seti için çalışmada kullanılan modele ait parametreler
(Parameters of the model used in the study for the SCOPe 2.06 dataset)

	Giriş Katmanı	**Birinci Gizli Katman**	**İkinci Gizli Katman**	**Üçüncü Gizli Katman**	**Çıktı Katmanı**
Nöron Sayısı	7	32	64	128	1430
Parametre Sayısı	0	256	2112	8320	184470
Kayıp Fonksiyonu	Kategorik Çapraz Düzensizlik Fonksiyonu (Categorical Cross Entropy)				
Aktivasyon Fonksiyonu	ReLU Fonksiyonu				
Optimizasyon Algoritması	Adam Optimizasyon Algoritması				

Her bir veri seti için yapay sinir ağının eğitiminde kullanılan veri sayısı, veri setine ait sınıf sayısı ve parametre sayıları Tablo 5'te verilmiştir. SCOPe 2.06, SCOPe 2.07 ve SCOPe 2.08 veri setleri sırasıyla genişletilmiş veri setleridir. Veri sayısı ve sınıf sayısının artığı görülmektedir.

Tablo 2. Farklı veri setlerine ait veri, sınıf ve parametre sayısı
(Number of data, classes and parameters for different data sets)

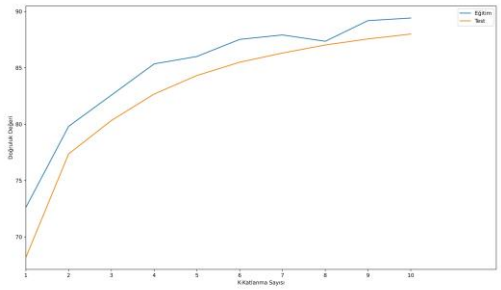
Veri Seti	**Veri Sayısı (Eğitim)**	**Sınıf Sayısı**	**Parametre Sayısı**
SCOPe 2.06	171028	1431	195287
SCOPe 2.07	193361	1457	198641
SCOPe 2.08	241395	1485	202253

SCOPe 2.06 veri seti ile yapılan çalışmada, katlama sayılarına ait doğruluk ve kayıp değerleri Tablo 3'te verilmiştir.

Tablo 3. SCOPe 2.06 veri seti için katlama numaralarına ait doğruluk ve kayıp değerleri
(Accuracy and loss values for fold numbers for SCOPe 2.06 dataset)

Katlama Sayısı	**Doğruluk Değeri**	**Kayıp Değeri**
1	72.6189	1.0517
2	79.8059	0.7604
3	82.5768	0.6191
4	85.3476	0.5192
5	86.0025	0.5192
6	87.5209	0.4938
7	87.9133	0.4472
8	87.3526	0.4274
9	89.1822	0.4256
10	89.4155	0.3754
Ortalama	84.7737	0.5685

SCOPe 2.06 veri seti ile katlama sayısı 10 alınarak gerçekleştirilen deneyde ortalama doğruluk değeri 84.77 elde edilmiştir. Literatürde yer alan diğer çalışmalara kıyasla iyi bir sonuç elde edilmiştir. SCOPe 2.06 veri setine ait katlama sayılarına göre eğitim ve test doğruluk değerlerinin grafiksel gösterimi Şekil 7'de verilmiştir.



Şekil 7. SCOPe 2.06 veri seti ile yapılan deneysel çalışmanın grafiksel gösterimi
(Graphical representation of the experimental work with the SCOPe 2.06 dataset)

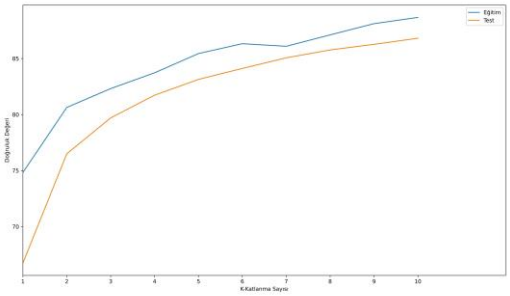
Şekil 7'de görüldüğü üzere katlama sayılarının her biri için eğitim doğruluk değerleri test doğruluk değerlerinden yüksek elde edilmiştir. SCOPe 2.07 veri seti ile yapılan çalışmada, katlama sayılarına ait doğruluk ve kayıp değerleri Tablo 4'te verilmiştir.

Tablo 4. SCOPe 2.07 veri seti için katlama numaralarına ait doğruluk ve kayıp değerleri
(Accuracy and loss values for fold numbers for SCOPe 2.07 dataset)

Katlama Sayısı	Doğruluk Değeri	Kayıp Değeri
1	74.8298	1.0474
2	80.6320	0.7370
3	82.314	0.6558
4	83.7237	0.5792
5	85.4432	0.5496
6	86.3229	0.4906
7	86.0912	0.4863
8	87.1085	0.4560
9	88.1113	0.4210
10	88.6689	0.4229
Ortalama	84.3247	0.5846

SCOPe 2.07 veri seti ile katlama sayısı 10 alınarak gerçekleştirilen deneyde ortalama doğruluk değeri 84.32 elde edilmiştir. Literatürde yer alan diğer çalışmalara kıyasla iyi bir sonuç elde edilmiştir. Ancak SCOPe 2.06 veri seti ile kıyaslandığında daha düşük doğruluk değeri elde edilmiştir. SCOPe 2.07 veri setine ait katlama sayılarına göre eğitim ve test doğruluk değerlerinin grafiksel gösterimi Şekil 8'de verilmiştir.

Şekil 8 incelendiğinde doğruluk değerlerinin katlama sayılarına göre sapmalar içerdiği görülse de değerlerin başarılı olduğu söylenebilmektedir. SCOPe 2.08 veri seti ile yapılan çalışmada, katlama sayılarına ait doğruluk ve kayıp değerleri Tablo 5'te verilmiştir. SCOPe 2.08 veri setine ait katlama sayılarına göre eğitim ve test doğruluk değerlerinin grafiksel gösterimi Şekil 9'da verilmiştir.

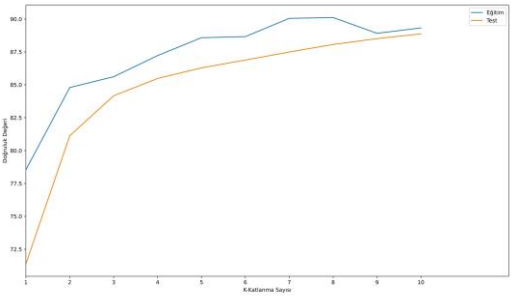


Şekil 8. SCOPe 2.07 veri seti ile yapılan deneysel çalışmanın grafiksel gösterimi
(Graphical representation of the experimental work with the SCOPe 2.07 dataset)

Tablo 5. SCOPe 2.08 veri seti için katlama numaralarına ait doğruluk ve kayıp değerleri
(Accuracy and loss values for fold numbers for SCOPe 2.08 dataset)

Katlama Sayısı	Doğruluk Değeri	Kayıp Değeri
1	78.5391	0.8482
2	84.7904	0.5661
3	85.6140	0.5088
4	87.2205	0.4507
5	88.5805	0.4208
6	88.6173	0.4103
7	90.0507	0.3539
8	90.1116	0.3628
9	88.9140	0.3915
10	89.3228	0.3668
Ortalama	87.1806	0.4680

SCOPe 2.08 veri seti ile gerçekleştirilen deneyde ortalama doğruluk değeri 87.18 elde edilmiştir. Çalışmamızda kullanılan diğer veri setleri ile kıyaslandığında en iyi sonuç SCOPe 2.08 veri seti ile elde edilmiştir.



Şekil 9. SCOPe 2.08 veri seti ile yapılan deneysel çalışmanın grafiksel gösterimi
(Graphical representation of the experimental work with the SCOPe 2.08 dataset)

Şekil 9'da yer alan grafik için de katlama sayıları aralarında aşağı ya da yukarı yönlü sapmalar olduğu görülmektedir. Ancak eğitim ve test doğruluk değerleri arasındaki sapmalar doğruluk değerleri ile değerlendirildiğinde uygun olduğu söylenebilmektedir. Tüm veri setlerine ait ortalama kayıp ve doğruluk değerleri varyansları ile hem eğitim hem de test sonuçları için Tablo 6 ve Tablo 7'de verilmiştir.

Tablo 6. Veri setlerine ait eğitim sonuçları
(Train results for data sets)

Veri Seti	Ortalama Kayıp Değeri	Ortalama Doğruluk Değeri	Varyans
SCOPe 2.06	0.5685	84.7737	± 4.9408
SCOPe 2.07	0.5846	84.3247	± 3.9687
SCOPe 2.08	0.4680	87.1806	± 3.3379

Veri setlerinin veri ve sınıf sayısı bakımından genişletilmesinin deney sonuçlarında olumlu yönde katkı sağlayacağı öngörülmüştür. Eğitim sonuçları incelendiğinde SCOPe 2.06 veri setine kıyasla daha güncel bir versiyon olan SCOPe 2.07 veri setinin, doğruluk değeri düşük çıkmıştır. Varyans değerleri ile karşılaştırma yapıldığında SCOPe 2.07 veri setinin SCOPe 2.06 veri setinden daha tutarlı olduğu söylenebilmektedir. SCOPe 2.06 ve SCOPe 2.07 veri setleri ile elde edilen sonuçlardaki varyans farkı, ortalama doğruluk değerleri karşılaştırıldığında kabul edilebilir görülmüştür.

Model üzerinde eğitim işlemi gerçekleştirilirken çalışmada kullanılan üç veri seti için de her katlamaya ait döngü sayısı (epoch) 300 olarak alınmıştır. Deneysel açıdan karşılaştırılabilir olması açısından döngü sayısı, döngü boyutu, katman sayısı, model yapısı gibi özellikler her veri seti için eşit alınmıştır. Bu doğrultuda en başarılı eğitim sonucu 87.18 doğruluk değeri ile SCOPe 2.08 veri setine ait olmuştur.

Tablo 7. Veri setlerine ait test sonuçları
(Test results for data sets)

Veri Seti	Ortalama Kayıp Değeri	Ortalama Doğruluk Değeri	Varyans
SCOPe 2.06	0.5947	83.3690	± 5.9900
SCOPe 2.07	0.5593	83.4462	± 5.6564
SCOPe 2.08	0.4912	86.8238	± 4.1862

Veri setlerine ait test doğruluk değerleri incelendiğinde, veri setinin iyileştirilmesi ile doğruluk değerlerinin iyileşmesi arasında pozitif ilişki olduğu görülmektedir. Çalışmada kullanılan her üç veri seti için de literatürde yer alan çalışmalara kıyasla daha iyi sonuçlar elde edilmiştir.

Bu çalışmada en iyi sonuç hem eğitim hem de test sonuçları için sırasıyla 87.18 ve 86.32 değerleri ile SCOPe 2.08 veri setine ait olmuştur.

Elde edilen benzerlik değerleri sonuçları kullanılarak eğitilen, yalnızca yapay sinir ağları kullanılan çalışmamız, literatürde yer alan birden fazla yöntemin birleştirildiği çalışmalara göre daha yüksek doğruluğa sahip sonuç elde etmiştir. Birden fazla yöntemin birleştirildiği çalışmalarda sistem performansı negatif olarak etkilenebilmektedir. Verilerde katlanma türü sayısının artışı protein katlanması sınıflandırması için ise pozitif yönde katkı sağlamaktadır.

Scop ve ScopE veri tabanlarında bulunan veri setleri kullanılarak yapılan çalışmalarda kullanılan yöntem ve bu yöntemlerden elde edilen doğruluk değerleri Tablo 8'de verilmiştir.

Tablo 8. Protein katlanması tanıma için literatürde yer alan çalışmaların karşılaştırması
(Comparison of studies in the literature for protein folding recognition)

Çalışma	Kullanılan Yöntem	Test Doğruluk Değeri
Levitt [1]	Destek Vektör Makinesi	%63
Suddha [2]	Destek Vektör Makinesi ve Yapay Sinir Ağları birleştirilmiş	%83
Xu [39]	ResNet	%57.7
Liu [15]	Destek Vektör Makinesi	%78
Liu [40]	Evrişimli Sinir Ağı (CNN)	%72.55
Morcillo [41]	Evrişimli Sinir Ağı (CNN), Kapılı Yinelemeli Üniteler (GRU) ve Karar Ağaçları birleştirilmiş	%76.3
Bu çalışma	Yapay Sinir Ağları	%86.32

Literatürde yer alan çalışmalar incelendiğinde protein katlanması tanıma işlemi için Destek Vektör Makinesi yönteminin sıklıkla kullanıldığı ve bu yöntemin yapay sinir ağı yöntemiyle birleştirildiği görülmüştür. SCOPe veri tabanlarındaki veri setlerinin güncellenmesiyle birlikte, proteinlere ait aile ve süper aile bilgisi, katlanma türü sayısının artması gibi etkenler doğruluk değerlerini artırmıştır. Yapay sinir ağlarının diğer yöntemlerle birleştirildiği çalışmalarda doğruluk değerlerinin önemli ölçüde artığı görülmüştür. Destek vektör makinesi ile yapay sinir ağları yöntemlerini birleştirilen çalışma [2], doğruluk değerleri karşılaştırmasına göre modelimize en yakın değeri elde etmiştir. Ancak iki farklı yöntemin kullanıldığı (destek vektör makinesi ve yapay sinir ağları) çalışmaya kıyasla yalnızca yapay sinir ağı kullanılan yaklaşımımızın performans açısından üstünlük sağladığı söylenebilir.

Çalışmamızda SCOPe veri tabanlarındaki 2.06, 2.07 ve 2.08 veri setlerinde bulunan proteinler ile ilgili özelliklerin tamamı kullanılmamıştır. Bilgi kazanımı hesaplanarak, en yüksek değere sahip DM, SP ve CL özellikleri kullanılmıştır. Bu sayede çalışmamızda kullanılan modelin performansında olumlu etki görülmüştür. Bu çalışmada elde edilen en yüksek test doğruluk değeri SCOPe 2.08 veri seti ile %86.82 olmuştur. Şablon tabanlı yöntemler kullanılarak elde edilen benzerlik değerleri sonuçları kullanılarak eğitilen, yalnızca yapay sinir ağları kullanılan çalışmamız, literatürde yer alan birden fazla yöntemin birleştirildiği çalışmalara göre daha yüksek doğruluğa sahip sonuç elde etmiştir. Birden fazla yöntemin birleştirildiği çalışmalarda sistem performansı negatif olarak etkilenebilmektedir. Verilerde katlanma türü

sayısının artışı protein katlanması sınıflandırması için ise pozitif yönde katkı sağlamaktadır.

5. TARTIŞMA (CONCLUSION)

Protein katlanması tanıma işlemi, protein yapılarının tahmini için gereklidir. Bu çalışmada yeni bir yaklaşım önerilmiştir. Literatür çalışmaları incelendiğinde protein katlanması tanıma yöntemi için farklı bilim dallarından faydalanılarak moleküler biyoloji alan için önem taşıyan protein katlanması tespiti ya da protein katlanması tahmini sayesinde protein yapılarının belirlenmesi işleminin gerçekleştirildiği görülmüştür. Özellikle proteinlerin kimyasal yapılarından faydalanılarak yapılan analizler sonucu katlanma türü tahmini ve sınıflandırması için farklı puanlama yöntemleri oluşturulmuş ve sınıflandırma işleminin iyileştirilmesi amacıyla protein veri bankalarında paylaşılmıştır. Bu çalışmada da protein dizilerinden, proteinlerin kimyasal bağ yapılarından elde edilen puanlar, yapay sinir ağı yöntemi ile birleştirilerek performansın iyileştirilmesi sağlanmıştır. Sunulan yeni yaklaşım, üç farklı veri seti açısından kıyaslanmıştır. Protein katlanması tanıma işlemi için literatürde yer alan diğer tahmin edici yöntemlerden, destek vektör makinesi, Bayes yöntemi, K en yakın komşu yöntemi gibi yaygın kullanılan yöntemlerden önemli ölçüde daha iyi performans gösteriyor.

TEŞEKKÜR (ACKNOWLEDGEMENTS)

Yazarlar, TÜBİTAK ULAKBİM, Yüksek Başarım ve Grid Hesaplama Merkezine (TRUBA), Bursa Teknik Üniversitesi Yüksek Performanslı Hesaplama Laboratuvarına teşekkürlerini sunar.

KAYNAKLAR (REFERENCES)

[1] M. Levitt, C. Chothia, *"Structural patterns in globular proteins."* Nature, 261(5561), 552-558, 1976.

[2] P. Sudha, D. Ramyachitra, P. Manikandan, *"Enhanced artificial neural network for protein fold recognition and structural class prediction"*, Gene Reports, 12, 261-275, 2018.

[3] J. S. Butler, S. N. Loh, *"Folding and misfolding mechanisms of the p53 DNA binding domain at physiological temperature"*, Protein science, 15(11), 2457-2465, 2006.

[4] Y. Kaya, R. Tekin, *"Epileptik nöbetlerin tespiti için aşırı öğrenme makinesi tabanlı uzman bir sistem"*, Bilişim Teknolojileri Dergisi, 5(2), 33-40, 2012.

[5] A. Haltaş, A. Alkan, *"Medline veritabanı üzerinde bulunan tıbbi dokümanların kanser türlerine göre otomatik sınıflandırılması*", Bilişim Teknolojileri Dergisi, 9(2), 181, 2016.

[6] G. Akgül, A.A. Çelik, Z.E. Aydın, Z. K. Öztürk, *"Hipotiroidi Hastalığı Teşhisinde Sınıflandırma Algoritmalarının Kullanımı*", Bilişim Teknolojileri Dergisi, 13(3), 255-268, 2020.

[7] A. Şenol, Y. Canbay, M. Kaya, "*Makine Öğrenmesi Yaklaşımlarını Kullanarak Salgınları Erken Evrede Tespit Etme Alanındaki Eğilimler*", Bilişim Teknolojileri Dergisi, 14(4), 2021.

[8] M. AlQuraishi, "*Machine learning in protein structure prediction*", Current opinion in chemical biology, 65, 1-8, 2021.

[9] C. Ekenna, S. Thomas, N.M. Amato, "*Adaptive local learning in sampling based motion planning for protein Folding*", BMC systems biology, 10(2), 165-179, 2016.

[10] J. Zhu, H. Zhang, S.C. Li, C. Wang, L. Kong, S. Sun, D. Bu, "*Improving protein fold recognition by extracting fold-specific features from predicted residue–residue contacts*" , Bioinformatics, 33(23), 3749-3757, 2017.

[11] K. Yan, J. Wen, J. X. Liu, Y. Xu, B. Liu, "*Protein fold recognition by combining support vector machines and pairwise sequence similarity scores*", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 18(5), 2008-2016, 2020.

[12] Y. Yang, E. Faraggi, H. Zhao, Y. Zhou, "*Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates*", Bioinformatics, 27(15), 2076-2082, 2011.

[13] S. Makigaki, T. Ishida, "Improvement of template-based protein structure prediction by using chimera alignment", **In Proceedings of the 2018 8th International Conference on Bioscience, Biochemistry and Bioinformatics,** Tokyo, Japonya, 32-37, Ocak 2018.

[14] J. Söding, A. Biegert, A.N. Lupas, "*The HHpred interactive server for protein homology detection and structure prediction*" , Nucleic acids research, 33(2), 244-248,2005.

[15] B. Liu,Y. Zhu, "ProtDec-LTR3.0: protein remote homology detection by incorporating profile-based features into learning to rank", Ieee Access, 7, 102499-102507, 2019.

[16] L. Wei, Q. Zou, "*Recent progress in machine learning-based methods for protein fold Recognition*", International journal of molecular sciences, 17(12), 2118, 2016.

[17] M. Corrales, P. Cusco, D. R. Usmanova, H.C. Chen, N.S. Bogatyreva, G.J. Filion, D.N. Ivankov, "*Machine learning: how much does it tell about protein folding rates?*", PloS one, 10(11), 2015.

[18] F. Noé, G. De Fabritiis, C. Clementi, "Machine learning for protein folding and Dynamics", **Current opinion in structural biology**, 60, 77-84, 2020.

[19] Internet: SCOPe: Structural Classification of Proteins — extended, https://scop.berkeley.edu, 15.05.2022.

[20] D. M. Halaby, A. Poupon, J. P. Mornon, "**The immunoglobulin fold family: sequence analysis and 3D structure comparisons**", Protein engineering, 12(7), 563-571,1999.

[21] T. J. Richmond, F. M. Richards, "*Packing of α-helices: Geometrical constraints and contact area*", Journal of molecular biology, 119(4), 537-555, 1978.

[22] T. J. P. Hubbard, T. L. Blundell, "**Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. Protein Engineering**", Design and Selection, 1(3), 159-171, 1987.

[23] J. Rozewicki,, S. Li,, K. M. Amada,D. M. Standley, K. Katoh, "MAFFT-DASH: integrated protein sequence and structural alignment", Nucleic acids research, 47(W1), W5-W10, 2019.

[24] V. Adar, Protein-ligand etkileşimleri, http://www.magum.hacettepe.edu.tr/MMKurs/KURS1Proteinligand.pdf, 24.06.2022.

[25] Y. N. Imai, Y. Inoue, L. Nakanishi, K. Kitaura, "*Cl–π interactions in protein–ligand complexes*", Protein Science, 17(7), 1129-1137, 2008.

[26] R. Rojas, **Neural Network A Systematic Introduction**, Springer, Heidelberg, Almanya, 1996.

[27] M. M. Yılmaz, **Periferik sinir defekt onarımında biyolojik kondüit modeli: de-epitelize insan amniyotik membranı ve adipoz kökenli mezenkimal kök hücre tabakası içeren sinir kondüit modelinin sinir iyileşmesine etkisinin değerlendirilmesi**, Uzmanlık Tezi, Hacettepe Üniversitesi, Tıp Fakültesi, 2020.

[28] J. Ma, J. Tang, "*A review for dynamics in neuron and neuronal network*", Nonlinear Dynamics, 89(3), 1569-1578, 2017.

[29] A. Eliasy, J. Przychodzen, "**The role of AI in capital structure to enhance corporate funding strategies**". Array, 6, 2020.

[30] I. H. Sarker, "*Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective*", SN Computer Science, 2(3), 1-16, 2021.

[31] V. Nair, G. E.Hinton, "*Rectified linear units improve restricted boltzmann machines*", Icml, 2010.

[32] X. Glorot, A. Bordes, Y. Bengio," Deep sparse rectifier neural networks", **In Proceedings of the fourteenth international conference on artificial intelligence and statistics**, Fort Lauderdale, A.B.D, 315-323, Nisan 2011.

[33] I. Goodfellow, Y. Bengio, A. Courville, "*Deep learning (adaptive computation and machine learning series*", Cambridge Massachusetts, 321-359, 2011.

[34] M. Bağ, **Derin öğrenme kullanarak IP üzerinden ses hizmeti veren şebekelerde sahtekarlığa yönelik çağrıların tespiti**, Yüksek Lisans Tezi, Ankara Üniversitesi, Fen Bilimleri Enstitüsü, 2019.

[35] Y.N.Fu'adah, N.K.C. Pratiwi, M.A. Pramudito, N. İbrahim,"Convolutional neural network (CNN) for automatic skin cancer classification system", **IOP Conf. Ser. Mater. Sci. Eng**., 982, 12005, 2020.

[36] G. Korkmaz, E. Eroğlu, "*Model karmaşıklığının kontrolü*", İktisadi ve İdari Yaklaşımlar Dergisi, 2(2), 146-162, 2020.

[37] B. Ö. Başer, M. Yangın, E.S. SARIDAŞ, "*Makine öğrenmesi teknikleriyle diyabet hastalığının sınıflandırılması*", Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 25(1), 112-120,2021.

[38] Z. Lyu, Y. Yu, B. Samali, M. Rashidi, M. Mohammadi, T.N. Nguyen, A. Nguyen, "*Back-propagation neural network optimized by K-fold cross-validation for prediction of torsional strength of reinforced Concrete beam"*, Materials, 15(4), 1477, 2022.

[39] J. Xu, *"Distance-based protein folding powered by deep learning*", Proceedings of the National Academy of Sciences, 116(34), 16856-16865, 2019.

[40] C. Li, B. Liu, "*MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks*", Briefings in Bioinformatics, 21(6), 2133–2141, 2020.

[41] A. Villegas-Morcillo, V. Sanchez, A.M. Gomez, "*FoldHSphere: deep hyperspherical embeddings for protein fold Recognition*", BMC bioinformatics, 22(1), 1-21, 2021.