ACTA INFOLOGICA 2022;6(1):43-52
dergipark.org.tr/acin

DOI: 10.26650/acin.1008075

RESEARCH ARTICLE

Derin Öğrenme (CNN, RNN, LSTM, GRU) Kullanarak Protein İkincil Yapı Tahmini

Using Deep Learning (CNN, RNN, LSTM, GRU) methods for the prediction of Protein Secondary Structure

Ezgi Çakmak[1], İhsan Hakan Selvi[2]

ÖZ

Protein, canlı organizmaların biyolojik süreçlerinde çok önemli bir role sahiptir. Proteinin işlevini bilmek, biyoloji ve tıp alanında gelecekteki çalışmalara büyük katkı sağlar. Proteinin fonksiyonunu anlamak için üç boyutlu yapısını anlamak önemlidir. Protein yapısını çözümlemek için X-ışını kristalografisi ve NMR gibi deneysel yöntemler kullanılmasına rağmen, sonuçların yetersiz olduğu kanıtlanmıştır. Bu nedenle, proteinlerin üç boyutlu yapısının tahmini, süreçlerdeki en önemli konulardan biri haline gelmektedir. Birincil yapı olarak bilinen amino asit dizisinden proteinin üç boyutlu şeklinin belirlenmesi zorlu olarak tanımlandığından, ikincil yapının tahmin edilmesi bu konuda önemli bir rol oynamaktadır. Literatürde protein ikincil yapısını tahmin etmek için makine öğrenmesi ve son zamanlarda derin öğrenme gibi birçok yöntem kullanılmıştır. Bu makale, yaygın olarak uygulanan dört derin öğrenme yöntemi olan CNN, RNN, LSTM ve GRU kullanılarak geliştirilen modellerin performanslarının bir karşılaştırmasını sağlamayı amaçlamaktadır. Bu modellerin eğitimi ve test edilmesi amacıyla CB513 veri seti kullanılmış, buna ek olarak doğruluk, f1 skoru, doğruluk ve kesinlik gibi performans değerlendirme ölçütleri uygulanmıştır. CNN, RNN, LSTM ve GRU modellerinin doğruluk oranları sırasıyla %82,54, %82,06, %81,1 ve %81,48'dir.

Anahtar kelimeler: Protein İkincil Yapı Tahmini, CNN, RNN, GRU

ABSTRACT

Proteins play a crucial function in the biological processes of living organisms. Knowing the function of the protein offers significant insight into future biological and medical research. Since a protein's shape determines its function, it is important to understand the protein's 3D structure. Although experimental methods such as X-ray crystallography and nuclear magnetic resonance (NMR) have been used to examine the shape of proteins, so far the results have been insufficient. As a result, predicting the 3D structure of proteins is crucial. Determining the 3D structure of a protein from its primary structure is challenging. Therefore, predicting the protein secondary structure becomes important for studying its structure and function. Many emerging methods, including machine learning, as well as deep learning, have been used to predict the secondary structure of proteins and comprise a crucial part of Structural Bioinformatics. The goal of this study is to compare the results generated by predictive models that were created using the four most frequently utilized deep learning methods: convolutional neural networks (CNN), recurrent neural networks (RNN), long short term memory networks (LSTM), and gated recurrent units (GRU). The CB513 dataset was used to train and test these models, and performance evaluation metrics viz. accuracy, f1 score, recall, and precision were applied. The CNN, RNN, LSTM, and GRU models had an accuracy of 82.54%, 82.06%, 81.1%, and 81.48%, respectively.

Keywords: Protein Secondary Structure Prediction, CNN, RNN, GRU

[*] Bu çalışma, Doç. Dr. İhsan Hakan Selvi danışmanlığında hazırlanan "Derin Öğrenme Yöntemi ile Protein İkincil Yapı Tahmini" başlıklı yüksek lisans tezinden üretilmiştir.

[1] (Arş. Gör.), Sakarya Üniversitesi, Bilgisayar ve Bilişim Bilimleri Fakültesi, Bilişim Sistemleri Mühendisliği, Sakarya, Türkiye
[2] (Doç Dr.), Sakarya Üniversitesi, Bilgisayar ve Bilişim Bilimleri Fakültesi, Bilişim Sistemleri Mühendisliği, Sakarya, Türkiye

ORCID: E.Ç. 0000-0002-6970-8651; İ.H.S.0000-0002-8837-2137

Corresponding author:
Ezgi ÇAKMAK
Sakarya Üniversitesi, Bilgisayar ve Bilişim Bilimleri Fakültesi, Bilişim Sistemleri Mühendisliği, Sakarya, Türkiye
E-mail address: ezgicakmakk@gmail.com

Submitted: 11.10.2021
Revision Requested: 16.01.2022
Last Revision Received: 16.02.2022
Accepted: 14.03.2022
Published Online: 28.04.2022

Citation: Çakmak, E. & Selvi, I. H. (2022). Derin öğrenme (CNN, RNN, LSTM, GRU) kullanarak protein ikincil yapı tahmini. *Acta Infologica, 6*(1), 43-52. https://doi.org/10.26650/acin.1008075

1. GİRİŞ

Proteinler, canlı organizmaların temel bileşeni olup tüm canlı yapılarında bulunarak biyolojik süreçlere katılırlar (Allison, 2007). Proteinlerin temel yapısı olan amino asit dizilimlerinde meydana gelen katlanmalar proteinin üç boyutlu yapısını ve fonksiyonunu belirler (Dill & MacCallum, 2012). Bu sebeple, proteinlerin yapılarının bilinmesi, fonksiyonlarını ve proteinlerin rol aldığı biyolojik süreçleri anlamak için oldukça önemlidir (Liu & Hsu, 2005).

Proteinler birincil ikincil, üçüncül ve dördüncül yapı durumlarında bulunmaktadır. Proteinlerin yapı taşı olarak bilinen amino asitler arasında kurulan peptid bağları ile oluşan polipeptid zinciri birincil yapı olarak adlandırılmaktadır. Birincil yapı proteinin üç boyutlu yapısını ve işlevini belirleyen temel yapıdır. Birincil yapıda meydana gelen düzenli katlanmalar proteinin ikincil yapısını oluşturur. Temel ikincil yapı motifleri sarmal (Helix), tabaka (Sheet) ve döngü (Loop) olarak tanımlanmaktadır. Üçüncül yapı, ikincil yapı durumda meydana gelen etkileşimler sonucunda oluşan üç boyutlu yapıdır. Birden fazla polipeptid zincirinden oluşan bazı proteinlerin oluşturduğu yapı ise dördüncül yapı olarak bilinmektedir.

Proteinlerin üç boyutlu yapıları X-ışını kırınımı, Nükleer Manyetik Rezonans ve elektron kristalografisi gibi deneysel yöntemlerle çözümlenebilmektedir. Ancak bu yöntemlerle çözümlenebilen protein sayısı sınırlı olduğundan ve zaman aldığından (Moraes, Evans, Sanchez-Weatherby, Newstead, & Stewart, 2014) bilgisayarlı teknikler kullanılarak gerçekleştirilen protein yapı tahmini çalışmaları öne çıkmaktadır. Proteinlerin amino asit dizilimi kullanılarak ikincil yapısında meydana gelen katlanmaların tahmin edildiği ikincil yapı tahmin çalışmaları, proteinlerin yapı tahmininde kullanılan önemli bir aşamadır (Branden & Tooze, 2012).

Protein ikincil yapı tahmini çalışmaları ilk olarak 1970'li yıllarda başlamıştır. İstatiksel metotlar kullanılarak gerçekleştirilen bu çalışmalar %50-60 tahmin başarısına sahiptir. Literatürde bu çalışmalar birinci nesil olarak tanımlanmaktadır (B. Rost & Sander, 2000). GOR (Garnier, Osguthorpe, & Robson, 1978) ve Chou-Fasman (Chou & Fasman, 1974) algoritmaları birincil nesil yöntemler arasında yer almaktadır. 1980-1990 yılları arasında, tek bir amino asit kalıntısı yerine kayan pencere yöntemi kullanılmaya başlanmıştır. Bu dönemde, yapay sinir ağları, en yakın komşu algoritmaları ve istatiksel bilgi kullanılarak geliştirilen yöntemler kullanılmıştır. İkincil nesil olarak değerlendirilen bu yöntemlerin başarı oranları ise %70'leri aşamamıştır.

Deneysel yöntemlerle çözümlenen protein yapı bilgisinin zaman içerisinde artması ile birlikte, verilerden öğrenen makine öğrenmesi yöntemleri tahmin çalışmalarında sıklıkla kullanılmıştır. Protein ikincil yapı tahmin çalışmalarında sıklıkla kullanılan makine öğrenmesi yaklaşımları arasında yapay sinir ağları (Burkhard Rost & Sander, 1994),(Kneller, Cohen, & Langridge, 1990), k-en yakın komşu (Yi & Lander, 1993),(Salamov & Solovyev, 1995), destek vektör makineleri (Chen, Tian, Zou, Cai, & Mo, 2007),(Nguyen & Rajapakse, 2003) yer almaktadır. Son yıllarda, farklı problemleri öğrenmede başarılı olan derin öğrenme yöntemleri protein ikincil yapı tahmini çalışmalarında da ön plana çıkmıştır.

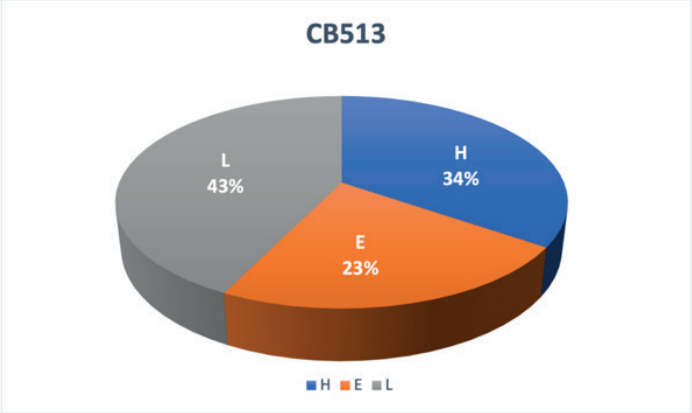
Evrişimli Sinir Ağları (CNN) ve Tekrarlayan Sinir Ağları (RNN) ikincil yapı tahmin çalışmalarında sıklıkla kullanılmaktadır. Baldi ve arkadaşları protein ikincil yapısını tahmin etmek için PSI-BLAST (Altschul et al., 1997) hizalaması ile elde ettikleri pozisyona özgü skorlama matrisi (PSSM) ve çift yönlü (bidirectional) tekrarlayan sinir ağı kullanarak %76 Q3 başarı oranı elde etmişlerdir (Baldi, Brunak, Frasconi, Soda, & Pollastri, 1999). Bu yöntem geliştirilerek Porter adı verilen bir web sunucusu oluşturulmuştur (Pollastri & McLysaght, 2005). Porter 4.0 protein ikincil yapı tahmininde %82,2 Q3 başarı oranına ulaşmıştır (Mirabello & Pollastri, 2013). Heffarnan ve arkadaşları tarafından geliştirilen yöntemde, çift yönlü tekrarlayan sinir ağı ve Uzun-Kısa Süreli Bellek (Long-Short Term Memory, LSTM) hücreleri kullanılmış ve %84,48 Q3 başarı elde etmişlerdir (Heffernan, Yang, Paliwal, & Zhou, 2017). Wang ve arkadaşları, SSREDNs (Secondary Structure Recurrent Encoder–Decoder Networks) adını verdikleri yöntemde çift yönlü geçitli yinelenen birim (Gated Rcurrent Unit, GRU) kullandıkları ağı CB513 veri seti ile test etmişler ve %82,9 Q3 başarı oranı elde etmişlerdir (Y. Wang, Mao, & Yi, 2017). Wang ve arkadaşları tarafından CNN ve LSTM katmanlarından oluşan yöntem %80,18 Q3 başarı oranına ulaşmıştır (J. Wang, Cheng, Zhao, & Lu, 2019). İkincil yapı tahmin çalışmalarında teorik limit %88 Q3 başarı oranı olarak belirtilmektedir (Rost, 2003).

Bu çalışmada, derin öğrenme yöntemlerinden CNN, RNN, LSTM ve GRU kullanılarak ağ yapıları oluşturulmuş ve protein ikincil yapı tahmini için CB513 veri seti kullanılarak ağlar eğitilmiş ve test edilmiştir. Ağların performansları başarı oranı, doğruluk, kesinlik ve F1 skoru hesaplanarak karşılaştırılmıştır.

2. MATERYAL VE YÖNTEM

2.1. Veri Seti

Cuff ve Barton tarafından oluşturulan CB513 veri seti (Cuff & Barton, 1999), protein ikincil yapı tahmin çalışmalarında sıklıkla kullanılmaktadır. Bu veri seti, toplamda 84119 amino asitten oluşan 513 protein içermektedir. Bu çalışmada kullanılan veri seti, CB513 veri setinde bulunan her bir amino asidin 539 öznitelikle temsil edildiği veri seti Aydın ve ark. çalışmasından (Aydin, Kaynar, & Görmez, 2018) elde edilmiştir. Veri setinde kullanılan öznitelik sayısının, her bir amino asit için hesaplanan 49 öznitelik ve kayan pencere yöntemi olarak adlandırılan hedef amino asidin ortada bulunduğu 11 amino asit seçilerek etrafındaki amino asitlerle etkileşimini ölçmek için kullanılan yöntem ile toplamda 49x11 olmak üzere hesaplandığı belirtilmiştir. Kullanılan veri seti, 513 proteinin her birinin eğitim ve test aşamalarında kullanılması için hazırlanmış olan yedi çapraz doğrulama kümesinden oluşmaktadır. Hazırlanan dört model belirtilen her çapraz doğrulama kümesi ile eğitilmiş ve test edilmiştir.



Şekil 2.1. CB513 Veri Seti İkincil Yapı Dağılımı

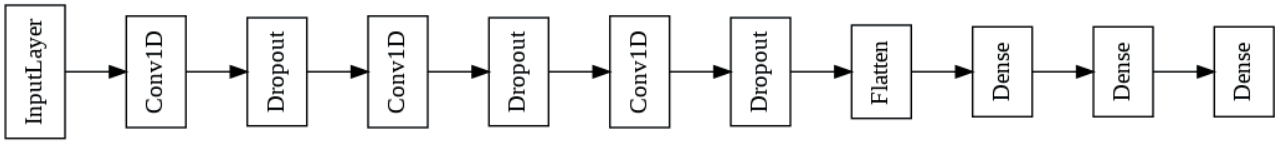
Veri setinde yer alan 84119 amino aside karşılık gelen sarmal (H), iplik (E) ve döngü (L) ikincil yapılarında bulunma sıklığı Şekil 2.1.'de gösterilmektedir. Bu ikincil yapı durumları veri setinde 0, 1 ve 2 değerleri ile temsil edilmektedir.

2.2. Derin Öğrenme Modellerin Geliştirilmesi

Bu çalışmada, protein ikincil yapısının tahmini için CNN, RNN, LSTM ve GRU olmak üzere farklı derin öğrenme modelleri kullanılarak dört farklı derin ağ oluşturulmuştur. Modeller, Google Colaboratory ("Colab," n.d.) platformu kullanılarak Python programlama diliyle hazırlanmıştır. Veriler, çalışma ortamına aktarıldıktan sonra program üzerinde *49x11* olmak üzere, 49 hedef amino asidin öznitelik sayısını ve 11 hedef amino asit etrafındaki seçilen amino asit sayısını (pencere boyutu) temsil etmektedir, yeniden boyutlandırılmış ve eğitim için hazır hale getirilmiştir. Amino aside karşılık gelen üç adet ikincil yapı durumu, sarmal (H), iplik (E) ve döngü (L), bulunduğundan üç sınıflı tahmin yapılmıştır. Her çapraz doğrulama eğitim setinin %10'u doğrulama seti olarak ayrılmıştır. Ağ katmanları ve parametreleri belirlenen modeller her çapraz doğrulama seti ile eğitilmiştir. Geliştirilen tüm modeller için aynı işlemler tekrarlanmıştır.

2.2.1 CNN Modeli

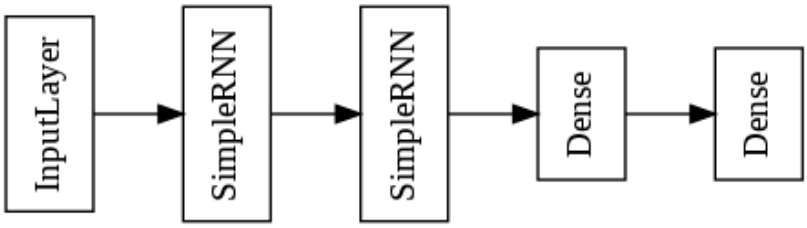
Evrişimli Sinir Ağları kullanılarak geliştirilen bu model, girdi katmanı, bir boyutlu konvolüsyon katmanları, seyreltme katmanları, düzleştirme katmanı ve yoğun katmanlardan oluşmaktadır. Konvolüsyon katmanlarında sırasıyla 128, 64 ve 32 filtre tanımlanmıştır. Kernel ise sırasıyla 5, 3 ve 3 olarak belirlenmiştir. Ağırlıkların düzenlenmesini sağlayan L2 regülasyonu bu katmanlarda 0,001 olarak seçilmiştir ve ReLu aktivasyon fonksiyonu kullanılmıştır. Seyreltme işlemi için 0,20 oranı belirlenmiştir. Çıktı katmanında sınıf sayısını temsil eden üç çıkış tanımlanmıştır ve sınıflandırma problemlerinde her sınıf için tahmin olasılığını hesaplayan softmax aktivasyon fonksiyonu kullanılmıştır. Şekil 2.2.'de CNN modelinin katman yapısı gösterilmektedir.



Şekil 2.2. CNN Modeli Katmanları

2.2.2. RNN Modeli

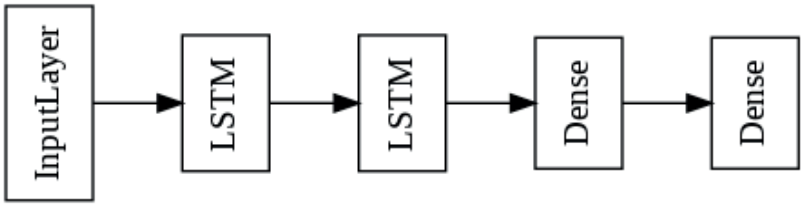
Yineleyen Sinir Ağları kullanılarak geliştirilin RNN modeli, iki RNN katmanı ile iki yoğun katmandan oluşmaktadır. RNN katmanlarında sırasıyla 64 ve 32 nöron tanımlanmıştır. RNN katmanları ve ilk yoğun katmanın aktivasyon fonksiyonu ReLu olarak belirlenmiştir. Çıktı katmanında, 3 nöron ve softmax aktivasyon fonksiyonu tanımlanmıştır. Modelin katman yapısı Şekil 2.3'te gösterilmektedir.



Şekil 2.3. RNN Modeli Katmanları

2.2.3. LSTM Modeli

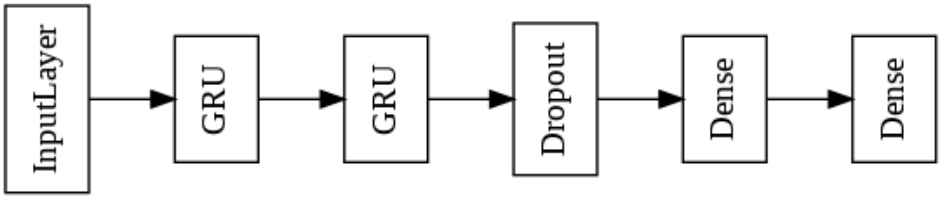
Kısa-Uzun Süreli Bellek kullanarak geliştirilen üçüncü model LSTM modeli olarak adlandırılıp iki LSTM katmanı ve iki yoğun katmandan oluşmaktadır. LSTM katmanlarında sırasıyla 128 ve 64 nöron yer almaktadır. LSTM katmanlarını aktivasyon fonksiyonu tanh, yoğun katmanın aktivasyon fonksiyonunda ise ReLu seçilmiştir. Çıktı katmanı diğer modellerde olduğu gibi 3 çıkıştan oluşmakta ve softmax aktivasyon fonksiyonu kullanılmaktadır. Şekil 2.4'te LSTM modelinin ağ katmanları gösterilmektedir.



Şekil 2.4. LSTM Modeli Katmanları

2.2.4. GRU Modeli

Bu çalışmada geliştirilen dördüncül modelde, geçitli yinelenen birim kullanılmıştır. Girdi katmanı, GRU katmanları, seyreltme ve yoğun katmanlardan oluşan modelin katman yapısı şekilde gösterilmektedir. GRU katmanlarında sırasıyla 100, 50, yoğun katmanda 50 nöron ile aktivasyon fonksiyonu olarak tanh tanımlanmıştır. GRU katmanlarının ardından 0,20 oranında seyrelme işlemi uygulanmaktadır. Çıktı katmanında, 3 nöron ile softmax aktivasyon fonksiyonu belirlenmiştir. Şekil 2.5.'te geliştirilen GRU modeli katmanları gösterilmektedir.



Şekil 2.5. GRU Modeli Katmanları

Modeller, 0,0001 öğrenme katsayısı ile ADAM optimizasyon algoritması ve 64 yığın boyutu kullanılarak eğitilmiştir.

2.3. Performans Değerlendirme Metrikleri

Bu çalışmada geliştirilen dört farklı derin öğrenme ağının performanslarını karşılaştırabilmek için, sınıflandırma problemlerinde kullanılan başarı oranı, kesinlik, duyarlılık ve F1 skoru değerleri hesaplanmıştır. Değerlendirme metriklerinin hesaplanması, Sklearn kütüphanesi kullanılarak gerçekleştirilmiştir.

2.3.1. Başarı Oranı

Q3 başarı oranı, üç sınıflı protein ikincil yapı tahmin çalışmalarında kullanılmakta ve sarmal (H), döngü (L) ve iplik (E) sınıflarının doğru tahmin edildiği örnek sayısının toplam örnek sayısına bölünmesi ile hesaplanmaktadır.

$$\boldsymbol{Başarı\ Oranı} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + FP + TN + FN}}$$

2.3.1. Duyarlılık

Çok sınıfllı tahmin çalışmalarında duyarlılık (recall) her sınıf için ayrı olarak, doğru tahmin edilen pozitif durumların (TP, True Pozitive), pozitif durumların toplamına bölünmesi ile hesaplanır. FN (False Vegative), yanlış tahmin edilen negatif durumları temsil etmektedir.

$$Duyarlılık = \frac{TP}{TP + FN}$$

2.3.1. Kesinlik

Kesinlik (precision), doğru tahmin edilen pozitif durumların (TP), pozitif tahmin edilen tüm durumlara bölünmesi ile elde edilir. Yanlış pozitif (FP, False Positive), yanlış tahmin edilen pozitif sınıfları temsil etmektedir.

$$Kesinlik = \frac{TP}{TP + FP}$$

2.3.1. F1 Skoru

Duyarlılık ve kesinlik değerlerinin harmonik ortalaması alınarak hesaplanan metrik F1 skoru olarak tanımlanmaktadır.

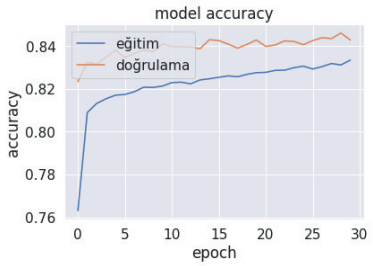
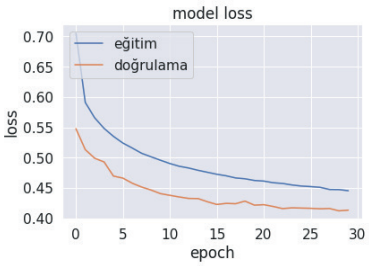
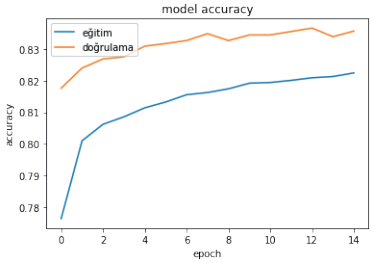
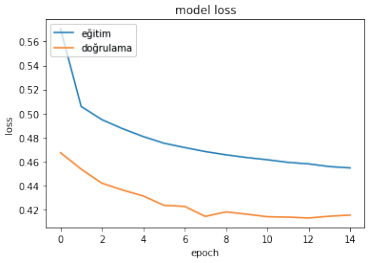
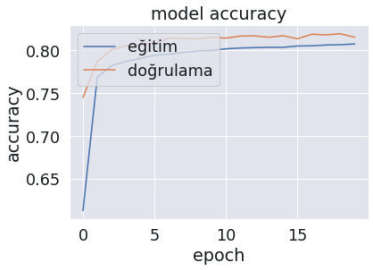
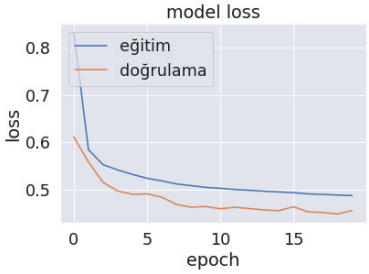
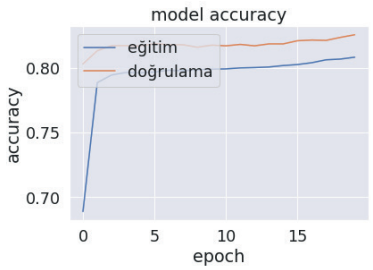
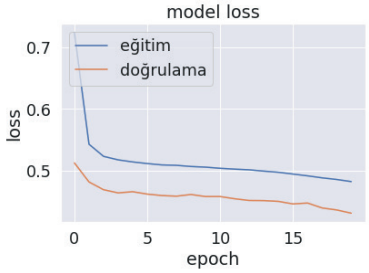
$$F1\ Skoru = 2 * \frac{Duyarlılık * Kesinlik}{Duyarlılık + Kesinlik}$$

3. ARAŞTIRMA BULGULARI

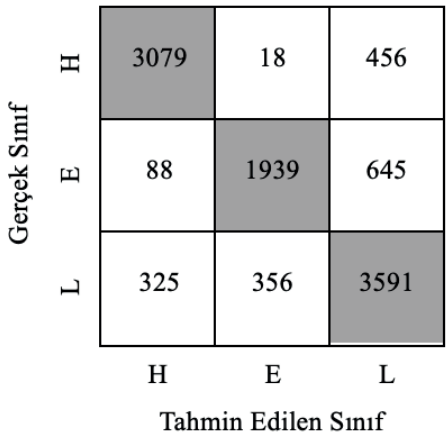
Bu çalışmada, CNN, RNN, LSTM ve GRU olmak üzere dört faklı derin öğrenme modeli geliştirilerek protein ikincil yapı tahmini yapılmıştır. Geliştirilen modellerin eğitimleri gerçekleştirildikten sonra, her çapraz doğrulama kümesi için test edilmiştir. Ağların eğitim ve test aşamaları Google Colaboratory platformunda GPU kullanılarak gerçekleştirilmiştir. RNN ağının GPU üzerinde eğitimi CPU üzerinde çalıştığından daha uzun sürdüğünden, tüm aşamaları CPU üzerinde yürütülmüştür.

Geliştirilen dört model, yedi çapraz doğrulama seti ile eğitilip test edildiğinden her model için 14 yakınsama grafiği elde edilmiş olup, her bir modelin birinci çapraz doğrulama seti ile eğitim aşamasında elde edilen yakınsama grafikleri örnek olarak Tablo 3.1.'de verilmiştir.

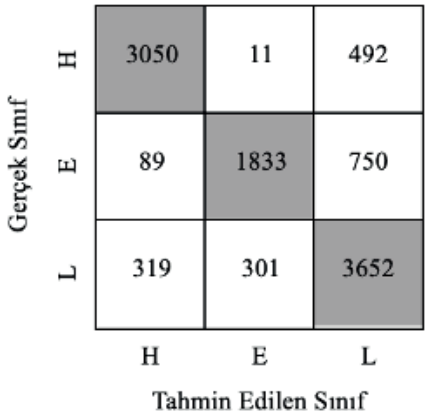
Tablo 3.1. *Birinci Çapraz Doğrulama Seti ile Modellerin Eğitimi Sonucunda Elde Edilen Başarı Oranı ve Kayıp Değeri Grafikleri*

Model	Başarı Oranı	Kayıp Değeri
CNN Modeli		
RNN Modeli		
LSTM Modeli		
GRU Modeli		

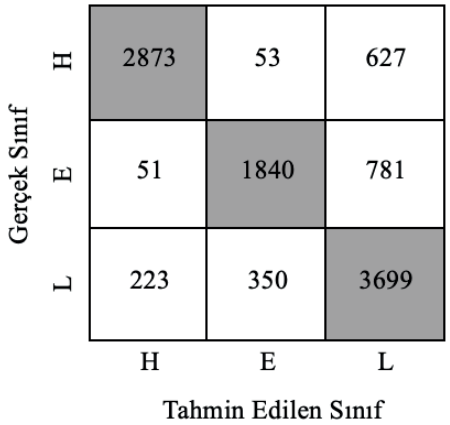
Eğitilen ağların her çapraz doğrulama seti ile test edilmesinden sonra performanslarını karşılaştırmak için kesinlik, duyarlılık, F1 skoru, Q3 başarı oranları ve karmaşıklık matrisleri hesaplanmıştır. Her bir modelin birinci çapraz doğrulama sonucunda hesaplanan karmaşıklık matrisleri örnek olarak Şekil 3.1., 3.2., 3.3. ve 3.4.'te gösterilmektedir.



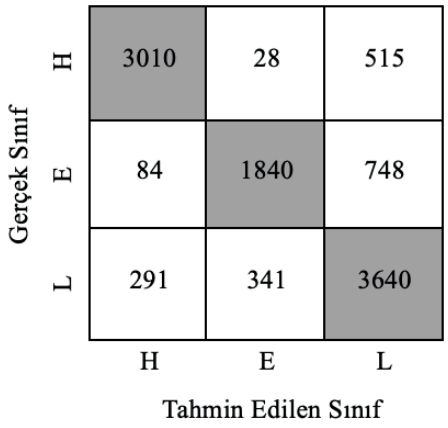
Şekil 3.1. CNN Modeli Birinci Çapraz Doğrulama Seti Karmaşıklık Matrisi



Şekil 3.2. RNN Modeli Birinci Çapraz Doğrulama Seti Karmaşıklık Matrisi



Şekil 3.3. LSTM Modeli Birinci Çapraz Doğrulama Seti Karmaşıklık Matrisi



Şekil 3.4. GRU Modeli Birinci Çapraz Doğrulama Seti Karmaşıklık Matrisi

Modellerin toplam eğitim süreleri, çapraz doğrulama setleri için hesaplanan değerlendirme metriklerinin ağırlıklı ortalamaları ve başarı oranlarının standart sapma değeri tablo 3.2.'de yer almaktadır.

Tablo 3.2. *Geliştirilen Modellerin Ortalama Değerlendirme Metrikleri*

	Eğitim Süresi	**Kesinlik**	**Duyarlılık**	**F1 skoru**	**Başarı oranı**	**Standart Sapma**
CNN	13 dk. 39 sn.	0,8267	0,82	0,82	0,8254	0,0100
RNN	50 dk. 52 sn.	0,8267	0,8133	0,82	0,8206	0,0081
LSTM	19 dk. 36 sn.	0,8167	0,80	0,81	0,8110	0,0087
GRU	16 dk. 18 sn.	0,8167	0,8034	0,81	0,8148	0,0087

Tablo 3.2.'de görüldüğü gibi geliştirilen CNN, RNN, LSTM ve GRU modelleri ile protein ikincil yapılarının tahmininde sırasıyla %82,54, %82,06, %81,1 ve %81,48 başarı elde edilmiştir. Modellerin duyarlılık değerleri 0,8167 ile 0,8267, duyarlılık değerleri 0,80 ile 0,82 arasında yer almaktadır. F1 skoru, CNN ve RNN modelleri için 0,82, LSTM ve GRU modelleri için ise 0,81 olarak hesaplanmıştır.

4. TARTIŞMA VE SONUÇLAR

Bu çalışmada, proteinlerin birincil yapı bilgisinden ikincil yapısının tahmini için CNN, RNN, LSTM ve GRU yöntemleri kullanılarak dört farklı derin öğrenme modeli geliştirilmiştir. Google Colaboratory ortamında hazırlanan derin ağlar, CB513 veri seti kullanılarak eğitilmiş ve test edilmiştir.

Test sonuçlarına göre, %82,54 doğruluk değeri ile CNN modeli başarı oranı en yüksek model olmuştur. LSTM modeli %81,1 doğruluk ile modeller arasında ikincil yapı tahmin başarısı en düşük model olmuştur. Ortalama F1 skorları değerlendirildiğinde, CNN ve RNN modellerinin diğer modellere göre %1 daha iyi sonuç elde ettiği görülmüştür. GPU üzerinde çalışan modeller arasında en hızlı eğitilen model CNN, en yavaş eğitilen modelin ise LSTM olduğu görülmüştür. RNN modelinin eğitimi CPU üzerinde çalıştırılmış ve 50 dk. 52 sn. de tamamlanmıştır.

Literatürde yer alan çalışmalar ile karşılaştırıldığında, RNN kullanılarak %76 başarı elde edilmiş olan (Baldi, Brunak, Frasconi, Soda, & Pollastri, 1999), CNN ve LSTM modellerinin hibrit olarak kullanılması sonucu %80,18 başarı edilen (J. Wang, Cheng, Zhao, & Lu, 2019) çalışmalardan daha iyi sonuç verdiği görülmektedir. Bu çalışmada RNN ile elde edilen %82,06'lık başarı oranının, RNN modeli kullanan Porter 4.0 test sonucunda %82,2 ile geçildiği görülmektedir (Mirabello & Pollastri, 2013). Mirabello ve Pollastri'nin çalışmasında kullanılan veri setinde aminoasit sayısının fazla olması ve veri setinin hacmi bu çalışmada kullanılan CB513 veri setinden daha büyük olduğu göz önüne alındığında başarının oldukça yakın olduğu görülmektedir. Rost'un çalışmasında belirttiği teorik başarı oranı olan %88 (Rost, 2003) baz alındığında ve literatürdeki çalışmaların başarı oranlarına bakıldığında %76 ile %85 arasında değiştiği gözlemlenmektedir. Bu çalışmada kullanılan tüm modellerin literatürde yer alan çalışmalar ile yakın sonuçlar verdiği görülmektedir.

Sonuç olarak, çalışmada geliştirilen modellerin tahmin başarılarının birbirine yakın olduğu görülmüştür. Modellerin eğitim sürelerinin literatürde yer alan diğer çalışmalardan oldukça kısa olması sebebiyle, tahmin hızının önemli olduğu durumlarda bu çalışmada geliştirilen derin ağ yöntemlerinin ve geliştirme ortamının kullanılabileceği görülmektedir.

Hakem Değerlendirmesi: Dış bağımsız.
Çıkar Çatışması: Yazarlar çıkar çatışması etmemişlerdir.
Finansal Destek: Yazarlar bu çalışma için finansal destek almadığını beyan etmemişlerdir.
Yazar Katkıları: Çalışma Konsepti/Tasarım- E.Ç.,İ.H.S.; Veri Toplama- E.Ç.,İ.H.S.; Veri Analizi/Yorumlama- E.Ç.,İ.H.S.; Yazı Taslağı- E.Ç.,İ.H.S.; İçeriğin Eleştirel İncelemesi- E.Ç.,İ.H.S.; Son Onay ve Sorumluluk- E.Ç.,İ.H.S.

Peer-review: Externally peer-reviewed.
Conflict of Interest: The authors have no conflict of interest to declare.
Grant Support: The authors declared that this study has received no financial support.
Author Contributions: Conception/Design of Study- E.Ç.,İ.H.S.; Data Acquisition- E.Ç.,İ.H.S.; Data Analysis/Interpretation- E.Ç.,İ.H.S.; Drafting Manuscript- E.Ç.,İ.H.S.; Critical Revision of Manuscript- E.Ç.,İ.H.S.; Final Approval and Accountability- E.Ç.,İ.H.S.

Kaynaklar/References

Allison, L. A. (2007). From gene to protein. In *Fundamental Molecular Biology* (1. Baskı). Blackwell Publishing.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Aydin, Z., Kaynar, O., & Görmez, Y. (2018). Comparison of NR and UniClust databases for protein secondary structure prediction. *2018 26th Signal Processing and Communications Applications Conference (SIU)*, 1–4. https://doi.org/10.1109/SIU.2018.8404285

Baldi, P., Brunak, S., Frasconi, P., Soda, G., & Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, *15*(11). https://doi.org/10.1093/bioinformatics/15.11.937

Branden, C. I., & Tooze, J. (2012). Introduction to Protein Structure. In *Introduction to Protein Structure*. https://doi.org/10.1201/9781136969898

Chen, C., Tian, Y., Zou, X., Cai, P., & Mo, J. (2007). Prediction of protein secondary structure content using support vector machine. *Talanta*. https://doi.org/10.1016/j.talanta.2006.09.015

Chou, P. Y., & Fasman, G. D. (1974). Prediction of Protein Conformation. *Biochemistry*. https://doi.org/10.1021/bi00699a002

Colab. (n.d.). Retrieved from https://research.google.com/colaboratory/intl/tr/faq.html

Cuff, J. A., & Barton, G. J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, *34*(4), 508–519. https://doi.org/https://doi.org/10.1002/(SICI)1097-0134(19990301)34:4<508::AID-PROT10>3.0.CO;2-4

Dill, K. A., & MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science*. https://doi.org/10.1126/science.1219021

Garnier, J., Osguthorpe, D. J., & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*. https://doi.org/10.1016/0022-2836(78)90297-8

Heffernan, R., Yang, Y., Paliwal, K., & Zhou, Y. (2017). Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btx218

Kneller, D. G., Cohen, F. E., & Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *Journal of Molecular Biology*. https://doi.org/10.1016/0022-2836(90)90154-E

Liu, H. L., & Hsu, J. P. (2005). Recent developments in structural proteomics for protein structure determination. *Proteomics*. https://doi.org/10.1002/pmic.200401104

Mirabello, C., & Pollastri, G. (2013). Porter, PaleAle 4.0: High-accuracy prediction of protein secondary structure and relative solvent accessibility. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btt344

Moraes, I., Evans, G., Sanchez-Weatherby, J., Newstead, S., & Stewart, P. D. S. (2014). Membrane protein structure determination—the next generation. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1838*(1), 78–87.

Nguyen, M. N., & Rajapakse, J. C. (2003). Multi-class support vector machines for protein secondary structure prediction. *Genome Informatics. International Conference on Genome Informatics*. https://doi.org/10.11234/gi1990.14.218

Pollastri, G., & McLysaght, A. (2005). Porter: A new, accurate server for protein secondary structure prediction. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bti203

Rost, B., & Sander, C. (2000). Third generation prediction of secondary structures. In *Webster, D. (Ed.), Protein Structure Prediction: Methods and Protocols* (pp. 71–95). https://doi.org/10.1385/1-59259-368-2:71

Rost, B. (2003). *Rising Accuracy of Protein Secondary Structure Prediction*. 207–249.

Rost, Burkhard, & Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Structure, Function, and Bioinformatics*. https://doi.org/10.1002/prot.340190108

Salamov, A. A., & Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *Journal of Molecular Biology*. https://doi.org/10.1006/jmbi.1994.0116

Wang, J., Cheng, J., Zhao, Z., & Lu, W. (2019). Protein Secondary Structure Prediction Using Ensemble of LSTM Neural Networks. *2019 2nd International Conference on Information Systems and Computer Aided Education, ICISCAE 2019*. https://doi.org/10.1109/ICISCAE48440.2019.221626

Wang, Y., Mao, H., & Yi, Z. (2017). Protein secondary structure prediction by using deep learning method. *Knowledge-Based Systems*, *118*, 115–123. https://doi.org/10.1016/j.knosys.2016.11.015

Yi, T. M., & Lander, E. S. (1993). Protein secondary structure prediction using nearest-neighbor methods. *Journal of Molecular Biology*. https://doi.org/10.1006/jmbi.1993.1464