T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DERİN ÖĞRENME YÖNTEMİ İLE PROTEİN İKİNCİL YAPI TAHMİNİ

YÜKSEK LİSANS TEZİ

Ezgi ÇAKMAK

Enstitü Anabilim Dalı	**:**	**BİLİŞİM SİSTEMLERİ MÜHENDİSLİĞİ**
Tez Danışmanı	**:**	**Doç. Dr. İhsan Hakan SELVİ**

Ağustos 2021

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DERİN ÖĞRENME YÖNTEMİ İLE PROTEİN İKİNCİL YAPI TAHMİNİ

YÜKSEK LİSANS TEZİ

Ezgi ÇAKMAK

Enstitü Anabilim Dalı : BİLİŞİM SİSTEMLERİ MÜHENDİSLİĞİ

Bu tez 19.08.2021 tarihinde aşağıdaki jüri tarafından oybirliği ile kabul edilmiştir.

Doç. Dr.	**Doç. Dr.**	**Doç. Dr.**
Ferdi SÖNMEZ	**İhsan Hakan SELVİ**	**Numan ÇELEBİ**
Jüri Başkanı	**Üye**	**Üye**

BEYAN

Tez içindeki tüm verilerin akademik kurallar çerçevesinde tarafımdan elde edildiğini, görsel ve yazılı tüm bilgi ve sonuçların akademik ve etik kurallara uygun şekilde sunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, tezde yer alan verilerin bu üniversite veya başka bir üniversitede herhangi bir tez çalışmasında kullanılmadığını beyan ederim.

Ezgi ÇAKMAK
19.08.2021

TEŞEKKÜR

Yüksek lisans eğitimim boyunca her konuda bilgi ve desteğini almaktan çekinmediğim, bu tezin hazırlanmasında yardımlarını esirgemeyen, değerli danışman hocam Doç. Dr. İhsan Hakan SELVİ'ye teşekkürlerimi sunarım.

Bu çalışma boyunca yardımlarını ve desteklerini esirgemeyen sevgili aileme teşekkür ederim.

İÇİNDEKİLER

SİMGELER VE KISALTMALAR LİSTESİ

CNN	: Convolutional Neural Network, Evrişimli Sinir Ağları
GRU	: Gated Recurrent Unit, Geçitli Yinelenen Birim
LSTM	: Long-Short Term Memory, Uzun-Kısa Süreli Bellek
NMR	: Nükleer Manyetik Rezonans
PDB	: Protein Data Bank
PİYT	: Protein İkincil Yapı Tahmini
RNN	: Recurrent Neural Network, Tekrarlayan Sinir Ağları
SVM	: Support Vector Machine, Destek Vektör Makinesi
YSA	: Yapay Sinir Ağları

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

ÖZET

Anahtar kelimeler: Derin Öğrenme, Protein İkincil Yapı, CNN, RNN

Protein yapısı tahmini, biyoinformatik alanındaki çalışmaların merkezi bir odak noktası olmuştur. Son yıllarda, proteinin yapısal bilgisini tahmin etmek için karmaşık makine öğrenmesi ve ardından derin öğrenme yöntemleri gibi birçok istatistiksel yöntem kullanılmıştır. Protein, canlı organizmaların önemli bir parçası olduğundan, protein yapısını ve işlevini anlamak ve değerlendirmek çok önemli hale gelmektedir. Proteinler, amino asit adı verilen yapı taşlarından oluşmaktadır. Protein yapısı büyük ölçüde birincil yapı olarak bilinen amino asit dizileri tarafından belirlenmesine rağmen, protein yapısını tek başına bu dizilerden tahmin etmek zordur. Dolayısıyla, sekanslardan protein ikincil yapı tahmini, proteinin üç boyutlu yapısının tahmini için önemli bir adımdır. Protein ikincil yapı tahmin çalışmalarında birçok yaklaşım kullanılmıştır. Bununla birlikte, günümüze kadar, literatürdeki tekniklerin hiçbiri tam olarak doğru bir sonuç verememiştir, bu da çalışmayı daha zorlu kılmaktadır. Bu tezde, CB513 veri setini kullanarak, derin öğrenme yaklaşımlarından, CNN, RNN, LSTM ve GRU kullanımına ilişkin karşılaştırmalı bir çalışma sağlanmaya çalışılmıştır. Çalışmada ayrıca her bir yaklaşımın performansı analiz edilmiş ve literatürdeki benzer çalışmalarla karşılaştırılmıştır. Bu çalışmada protein ikincil yapının tahmini için geliştirilen CNN, RNN, LSTM ve GRU modelleri sırasıyla %82,54, %81,06, %81,10, %81,48 başarı oranı elde etmiştir.

PROTEIN SECONDARY STRUCTURE PREDICTION USING DEEP LEARNING METHOD

SUMMARY

Keywords: Deep Learning, Protein Secondary Structure, CNN, RNN

Protein structure prediction has been a central focus of study in Bioinformatics. In the past few decades, many statistical methods, such as complex machine learning, followed by deep learning methods have been applied to estimate structural information of protein. Since protein is a significant part of living-organisms, understanding and assessing protein and its functions becomes crucial. Proteins are made by building block, called amino acid. Although protein structure is largely determined by amino acid sequences, known as primary structure, it is difficult to predict protein structure from those sequences alone. Thus, protein secondary structure prediction from the sequences is an important step for the estimation of protein three-dimensional structure. Many approaches have been employed onto protein secondary structure prediction studies. However, up to present days, none of the available techniques in literature is able to provide a fully accurate result, which makes the study more challenging. By using CB513 dataset, this thesis attempts to provide a comparative study of the use of deep learning approaches, CNN, RNN, LSTM and GRU. In the study, the performance of each approach was analyzed and compared with the similar studies in literature. The models, CNN, RNN, LSTM and GRU, developed for protein secondary structure prediction in this study achieved %82,54, %81,06, %81,10, %81,48 accuracy.

BÖLÜM 1. GİRİŞ

Proteinler, canlı organizmaların hemen her biyolojik sürecinde önemli rol alırlar. Proteinlerin fonksiyonlarını ve üç boyutlu yapılarını kendilerine özgü aminoasit dizilimleri belirler [1]. Farklı aminoasit dizilimine sahip proteinlerin fonksiyonları farklıdır [2]. Bu sebeple, proteinlerin amino asit sıralanışının bilinmesi biyolojik aktivitesinin açıklanmasında önemli bir aşamadır.

Proteinler yapılarına göre dört seviyeye ayrılırlar; birincil, ikincil, üçüncül ve dördüncül yapı. Aminoasitlerin peptid bağları bir araya gelerek oluşturduğu polipeptid zinciri birincil yapı olarak tanımlanmaktadır. Bu polipeptid zincirindeki herhangi bir değişiklik proteinin üç boyutlu yapısını ve dolayısıyla aktivitesini etkiler. İkincil yapı, polipeptid zincirinde görülen bölgesel katlanmalar sonucu oluşur, bu katlanmalar amino asitlerin karboksil ve amino grupları arasında kurulan hidrojen bağları ile meydana gelmektedir. Üçüncül yapı, proteinlerin üç boyutlu yapısını kazandığı, ikincil yapıda bulunan aminoasitlerin R grupları arasındaki etkileşimlerle oluşan yapıdır. Birden fazla polipeptid zincirinden oluşan proteinlerin bulunduğu yapı ise dördüncül yapı olarak tanımlanmaktadır.

Protein üç boyutlu yapısının bilinmesi, proteinin işlevini anlamanın yanı sıra genlerin işlevini anlama, protein katlanmalarındaki hatalardan kaynaklanan hastalıkların tespiti ve ilaç tasarımı için de oldukça önemlidir [1,3]. Proteinlerin üç boyutlu yapılarını belirlemek için X-ışını kırınımı, nükleer manyetik rezonans (NMR) ve elektron kristalografisi gibi deneysel teknikler kullanılmaktadır. Ancak laboratuvar çalışmaları ile bu yapıları belirlemek oldukça maliyetli ve zorlu olması yanı sıra her protein için bu teknikleri kullanmak mümkün değildir [3]. Proteinlerin üç boyutlu yapılarının, temelde onu oluşturan birincil yapıdan tahmini ise zor bir problem olarak

görülmektedir [4]. Bu sebeple, amino asit diziliminden (birincil yapı) ikincil yapı tahmini, protein yapısının ve dolayısıyla işlevinin anlaşılmasında önemli bir aşamadır.

Yıllar boyunca, protein ikincil yapısını tahmin etmek için çeşitli yöntemler kullanılmıştır. Protein veri bankasında (PDB) bulunan protein veri bilgisinin giderek artmasıyla beraber bilgisayarlı hesaplama teknikleri bu problemin çözümünde yaygın olarak kullanılmıştır. Yapay sinir ağları, destek vektör makineleri, genetik algoritmalar gibi makine öğrenmesi yaklaşımları bu araştırma alanında kullanılan yöntemler arasındadır. Son yıllarda yapılan çalışmalar, derin öğrenme modellerinin birçok farklı karmaşık problemde olduğu gibi protein ikincil yapı tahmin başarısını arttırdığı ortaya koymuştur.

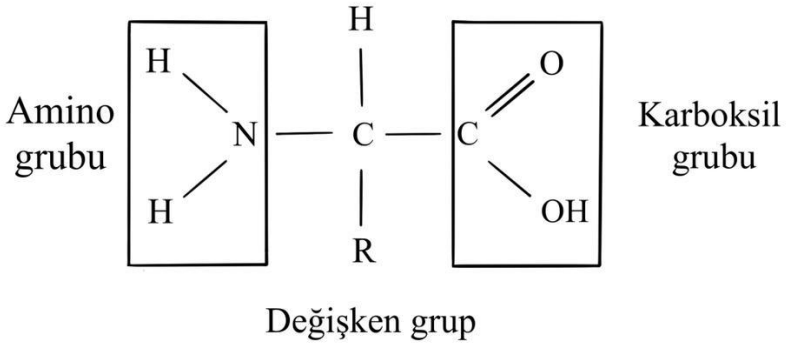
Bu tez çalışmasında, protein ikincil yapısının tahmini için CNN, RNN, LSTM VE GRU derin öğrenme modelleri ile çalışılmış ve performansları karşılaştırılmıştır. Eğitim ve test aşamalarında CB513 veri seti kullanılmıştır.

Bu tez çalışmasının ikinci bölümünde proteinlerin yapıları hakkında bilgi verilmiş ve protein ikincil yapı tahmin çalışmalarının tanıtıldığı kaynak araştırmasına yer verilmiştir. Üçüncü bölümde, çalışmada kullanılan derin öğrenme yöntemleri ile ilgili teknik bilgiler paylaşılmıştır. Dördüncü bölümde, bu çalışmada kullanılan çalışma ortamı, geliştirilen derin öğrenme yöntemleri ve çalışma adımları tanıtılmıştır. Beşinci bölümde, çalışmada kullanılan dört yöntemin eğitim ve test bulgularına yer verilmiştir. Altıncı bölümde, çalışmanın sonuçları paylaşılmıştır.

BÖLÜM 2. PROTEİNLER VE KAYNAK ARAŞTIRMASI

2.1. Proteinler

Proteinler, temelde amino asitlerin peptid bağı ile bir araya gelmesi ile oluşurlar. Amino asit dizilimindeki katlanmalar sonucunda ise proteinlerin üç boyutlu yapısı meydana gelir. Doğada proteinleri oluşturan 20 farklı amino asit türü mevcuttur. Amino asitler, merkez karbon atomuna (C_α) bağlı hidrojen atomu (H), amino grubu ($-NH_2$), karboksil grubu (-COOH) ve bir değişken yan zincirden (R) meydana gelirler. Şekil 2.1.'de amino asitlerin genel yapıları gösterilmektedir.



Şekil 2.1. Amino asitlerin genel gösterimi

Peptid zincirleri, bir amino asidin karboksil grubundaki hidroksil ile diğer amino asidin amino grubundaki hidrojen arasında peptid bağı kurulması ile oluşur ve amino asit sayısına göre isimlendirilirler; iki amino asit içeren zincir dipeptid, üç amino asit içeren zincir tripeptid, üçten fazla amino asidin birleşimi ile oluşan zincir ise polipeptid olarak adlandırılır. Zinciri oluşurken, solda kalan amino asidin bağ oluşmayan tarafında amino grubu bulunduğu için bu uç amino ucu ya da N-ucu olarak adlandırılırken, sağ uçtaki amino asidin karboksil grubu bağ oluşturmadığından bu uç

karboksil ucu ya da C-ucu olarak adlandırılır. Polipeptid zincirleri okunurken N-ucundan C-ucuna yani soldan sağa doğru okunurlar.

Proteinler, bir veya daha fazla polipeptid zincirinden oluşabilirler. Polipeptid zincirinde bulunan amino asitlere kalıntı (residue) adı verilir. Bunun sebebi, protein yapısına katılan amino asitleri, serbest haldeki amino asitlerden ayırabilmektir [5].

Amino asitler, üç harfli ve tek harfli kısaltmaları bulunmaktadır ancak protein yapısı gösterilirken genellikle tek harfli kısalmaları kullanılmaktadır. Proteinlerin yapısında bulunan amino asitlerin üç harfli ve tek harfli kısalmaları Şekil 2.2.'de gösterilmiştir.

Amino asit	**Kısaltma**		**Amino asit**	**Kısaltma**	
Glisin	Gly	G	Treonin	Thr	T
Alanin	Ala	A	Sistein	Cys	C
Valin	Val	V	Metiyonin	Met	M
Lösin	Leu	L	Asparajin	Asn	N
İzolösin	Ile	I	Glutamin	Gln	Q
Prolin	Pro	P	Aspartat	Asp	D
Fenilalanin	Phe	F	Glutamat	Glu	E
Tirozin	Tyr	Y	Lizin	Lys	K
Triptofan	Trp	W	Arjinin	Arg	R
Serin	Ser	S	Histidin	His	H

Şekil 2.2. Amino asitlerin kısaltmaları [6]

2.1.1. Protein yapısı

Proteinlerin yapısında dört seviye bulunmaktadır; birincil, ikincil, üçüncül ve dördüncül yapı. Her yapı, proteini oluşturan amino asit dizisinde kurulan bağ yapılarına bağlı olarak meydana gelir. Birincil yapıda meydana gelen katlanmalar ikincil yapıyı, ikincil yapıda bulunan zincirdeki etkileşimler üçüncül yapıyı ve üçüncül yapı da dördüncül yapıyı oluşturmaktadır.

2.1.1.1. Birincil yapı

Amino asitler arasında kurulan peptid bağları ile oluşan amino asit dizisi bir proteinin birincil yapısını oluşturur. Polipeptid zincirlerindeki merkezi amino asitlerin değişmeyen grupları arasında kurulan peptid bağları sebebiyle her zincirde aynı

olmasına rağmen farklılığı sağlayan amino asit kalıntılarının değişken R gruplarıdır. Birincil yapıyı oluşturan polipeptid zincirindeki amino asit türü, sayısı ve sırasındaki herhangi bir değişiklik farklı katlanmalara ve dolayısıyla farklı protein yapılarının oluşmasına neden olur. Birincil yapı proteinin üç boyutlu yapısını etkilediğinden dolaylı olarak proteinin fonksiyonunu da etkilemektedir [2].

2.1.1.2. İkincil yapı

Polipeptid zincirindeki amino asit kalıntılarının arasında kurulan hidrojen bağları, zincirde bazı düzenli ve düzensiz bölgesel katlanmalar meydana getirir. Proteinin ikincil yapısını oluşturan bu katlanmalar, iki temel motif oluşturur. Bu motifler, α-sarmal (α-helix) ve β-tabaka (β-sheet) olarak isimlendirilir. İki temel ikincil yapı dışındaki düzensiz katlanmalar ise rastgele döngü yapılarıdır (β-turn, loop). Bu motifler bölgesel olduklarından, bir proteinde birden fazla ikincil yapı bulunabilir.

α-sarmal: Birbirine dört amino asit kalıntısı uzaklıktaki kalıntılardan ilkinin -NH grubu ile diğer kalıntının -CO grubu arasında kurulan hidrojen bağı ile meydana gelen düzenli bükülmeler sonucu spiral şeklini alan polipeptid zinciri sarmal yapı olarak tanımlanır. Amino asit kalıntılarının R grupları, bu sarmal yapının dış tarafında kalır. α-sarmal en yaygın görülen protein ikincil yapıdır.

β-tabaka: Polipeptid zincirindeki amino asit kalıntıları arasında iki veya üç hidrojen bağının yan yana gelerek oluşturduğu tabakalardan meydana gelen akordeon şeklindeki yapıdır. Kırmalı tabaka olarak da adlandırılan β-tabaka, α-sarmal yapıdan sonra en sık rastlanan ikincil yapıdır.

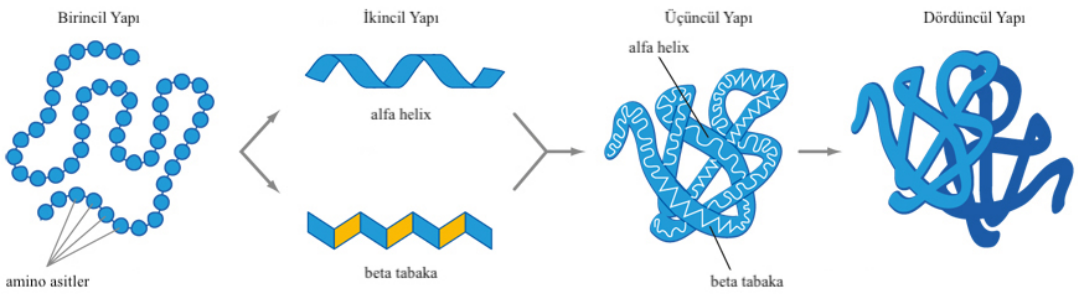
Döngü: Sarmal ve tabaka gibi düzenli bir tekrarı bulunmayan yapılardır ve genellikle sarmal ve tabaka yapıları arasında bulunarak bu yapıları birbirine bağlarlar.

2.1.1.3. Üçüncül yapı

Tek polipeptid zincirinden oluşan proteinlerde, ikincil yapı düzeyinde bulunan amino asit dizisinde yer alan kalıntıların R grupları arasında gerçekleşen etkileşimler sonucunda kazanılan üç boyutlu yapı proteinin üçüncül yapısını oluşturur. Hidrofobik yan zincire sahip amino asit kalıntıları proteinin iç bölgelerinde toplanırken, hidrofilik yan zincire sahip amino asit kalıntıları ise proteinin dış bölgelerinde yer alırlar. Van der Waals etkileşimler, disülfid bağları, iyonik bağlar üçüncül yapının oluşumunda etkili olan temel bağları oluşturmaktadır.

2.1.1.4. Dördüncül yapı

Bazı proteinler üçüncül yapıda bulunan birden fazla polipeptid zinciri içermektedir. Alt ünite olarak adlandırılan bu birimler arasında oluşan etkileşimler ve bağlar proteinin dördüncül yapısını oluşturur. Her protein üçüncül yapıda bulunan çok sayıda polipeptid zinciri içermediğinden dördüncül yapısı bulunmayabilir. Hemoglobin dördüncül yapıda bulunan proteinler arasında yer alır.



Şekil 2.3. Protein yapılarının illüstrasyonu [7]

Şekil 2.3.'te protein seviyelerinin illüstrasyonu yer almaktadır. Amino asitlerin oluşturdukları zincir birincil yapıyı, bu zincirdeki katlanmalar sonucu meydana gelen iki temel motif olan alfa helix ve beta tabaka ikincil yapıyı, ikincil yapı figürlerinin bir araya gelmesi üçüncül yapıyı ve son olarak birden fazla üçüncül yapının bir arada bulunduğu proteinlerin bulunduğu yapı ise dördüncül yapı olarak gösterilmektedir.

2.2. Protein Yapı Tahmini

Proteinlerin üç boyutlu yapıları deneysel yöntemlerle analiz edilebiliyor olmalarına rağmen bu yöntemler oldukça maliyetli oldukları gibi her protein için başarılı sonuç vermediklerinden, deneysel yöntemlerle çözümlenmiş protein yapılarından yola çıkarak yapısı belirlenmemiş proteinler için model geliştirmek protein yapı tahmini olarak tanımlanmaktadır. Ancak proteinin üç boyutlu yapısının tahmini zor bir problem olarak tanımlandığından, proteinlerin üç boyutlu yapısının tahmininde kullanılmak üzere ikincil yapı tahmini, çözücü erişilirlik sıklıkla çalışılan konular arasında yer almaktadır.

2.2.1. Protein İkincil Yapı Tahmini

Protein ikincil yapısının tahmini, proteinin birincil yapısından yani amino asit dizisinden her amino aside karşılık gelen ikincil yapı (α-sarmal, β-tabaka gibi) sınıfının tahmin çalışılmalarından oluşmaktadır. Kabsch ve Sander tarafından oluşturulan DSSP (Dictionary of Secondary Structure) adı verilen veri tabanı, PDB'de yer alan her protein için protein ikincil yapı sınıflarını hidrojen bağlanmalarına göre sekiz standart sınıfla temsil eder [8]. Protein ikincil yapı tahmininde bu sekiz sınıf kullanılabildiği gibi, genellikle sekiz sınıflı yapı üç ana sınıfa indirgenerek üç sınıflı yapı tahmini yapılmaktadır. Tablo 2.1.'de her bir ikincil yapı sınıfının harf karşılığı ve sekiz sınıfın üç sınıfa indirgemesinde kullanılan yöntem gösterilmektedir.

Tablo 2.1. Sekiz sınıflı ve üç sınıflı ikincil yapı gösterimi

DSSP Sekiz sınıflı ikincil yapı		Üç sınıflı ikincil yapı	
H	α-sarmal	H	Sarmal (Helix)
G	3_{10}-sarmal (3_{10} helix)		
I	5-sarmal (π-helix)		
E	extended strand, participates in β ladder	E	Beta İplik (β-Strand)
B	residue in isolated β-bridge		
T	hidrojen bağı dönüş	L	Döngü (Loop)
S	bend		
' ' veya -	coil		

2.3. Kaynak Araştırması

Protein yapısındaki aminoasit dizilimlerinde meydana gelen hidrojen bağlarının temelde sarmal ve tabaka adı verilen iki motif oluşturduğu, bu motiflerin ise coil veya turn olarak adlandırılan düzensiz motiflerle bağlandığı Pauling ve Corey tarafından yapılan çalışmalarda tahmin edilmiştir [9]–[12]. Bu çalışmaları ile Pauling 1954 yılında Nobel Kimya ödülünü almaya hak kazanmıştır.

1950'li yılların sonlarında Perutz ve doktora öğrencisi Kendrew, globular protein olan hemoglobin ve miyoglobinin üç boyutlu moleküler yapılarını X-ışını kırınımı yöntemi ile tanımlayarak bir proteinin yapısının ilk kez atomik seyide anlaşılmasını sağlamışlardır [13], [14]. 1962 yılında, Perutz MRC Moleküler Biyoloji Laboratuvarını kurmuştur. Aynı yıl, çalışmalarıyla yakaladıkları başarı sebebi ile Nobel Kimya ödülü almışlardır.

Rost ve Sander, protein ikincil yapı tahmini yöntemlerini üç nesil olarak incelenmiştir [15]. 60'lı ve 70'li yıllarda, amino asit kalıntıları ile protein ikincil yapısında bulunan motifler (α-sarmalı, β-tabaka) arasındaki ilişkinin istatiksel metotlarla analiz edildiği yöntemler birinci nesil olarak tanımlanmıştır. Bu yöntemle yapılan çalışmalardaki tahmin başarı oranı, o dönemde deneysel olarak gözlemlenebilen protein sayısına bağlı olarak, %50-60 arasında seyretmiştir.

Chou-Fasman tarafından deneysel olarak analiz edilmiş, bilinen proteinler kullanılarak her bir amino asit için yapısal elementlerde bulunma sıklığı hesaplanmış ve sonuçlara bakılarak herhangi bir amino asit dizisi için ikincil yapı tahmini geliştirilmiştir [16], [17]. Birinci nesil yöntemlerden olan bu metot literatürde, Chou-Fasman algoritması olarak yer almaktadır.

Garnier ve arkadaşları tarafından geliştirilen GOR, Chou-Fasman algoritmasında olduğu gibi olasılıklara dayanan ancak yalnızca tek amino asit için ikincil yapı oluşturma olasılığını incelemek yerine 20 aminoasit içerisinde 17 boyutlu pencerelerle

komşu amino asit kalıntılarının bulunduğu ikincil yapı durumunu da hesaplayan koşullu olasılık kullanılan bir metottur [18].

İkinci nesil yöntemler, birinci nesle kıyasla daha büyük veri setleriyle çalışılan, merkez aminoasit için mevcut ikincil yapıyı temel alan kayan pencere yönteminin ağırlıklı olarak kullanıldığı, 1990'lı yılların başına kadar çalışılan yöntemlerden oluşmaktadır [15]. En sık uygulanan ikincil nesil yöntemler arasında en yakın komşu algoritmaları [19], yapay sinir ağları [20] ve istatiksel bilgiye [21] dayanan çalışmalar yer almaktadır.

Gibrat ve arkadaşları tarafından GOR algoritmasından geliştirilen, istatiksel bilgi kullanılan GOR3 ikinci nesil yaklaşımların başında gelen çalışmalardan olmuştur. Bu çalışmada araştırmacılar, 68 proteinde ikincil yapı durumlarını tahmin ettiklerini ve daha önceki versiyonuna göre %7 başarı artışı sağlayarak %63 doğru tahmin başarısına ulaştığını belirtmişlerdir [21].

Ptitsyn ve Finkelstein stereokimyasal tahmin kurallarını kullanarak geliştirdikleri algoritma ile helix için %74 doğru tahmin sağlarken, β-strand için %58 doğruluk oranına ulaşabilmiştir [22].

Rost ve Sander, birinci nesilden Chou-Fasman, Lim ve GOR1 ile ikinci nesilden olan ALB ve Scheider algoritmalarını aynı veri setini kullanarak karşılaştırmış ve ikinci nesil algoritmaların %10 daha başarılı olduğunu belirtmişlerdir [15]. Ancak yine de ikinci nesil yöntemlerin başarı oranı %70'leri aşamamıştır. Başarı oranlarının istenilen seviyeye ulaşamaması ve beta tabaka durumunun sarmal yapıya göre daha düşük tahmin seviyesinin olması gibi birçok sebep araştırmacıları üçüncü nesil yöntemlere yönlendirmiştir.

90'lı yıllardan günümüze kadar olan süreçte, online veri tabanlarındaki çözümlenmiş protein verilerinin artması ve teknolojik gelişmelerle birlikte ikincil yapı tahmin çalışmalarında makine öğrenmesi teknikleri üzerine yoğunlaşan üçüncü nesil yöntemler kullanılmıştır. Protein ikincil yapı tahmininde en sık kullanılan makine

öğrenmesi tekniklerini yapay sinir ağları [23]–[25], saklı Markov modeli [26]–[28], destek vektör makineleri [29]–[31] oluşturmaktadır. Üçüncü nesil çalışmalar ile beraber ikincil yapı tahmin başarısı %70'leri aşmıştır.

Biyoinformatik alanında iki veya daha fazla dizinin benzer bölgelerini belirlemek için çoklu dizi hizalama yöntemleri kullanılmaktadır. Protein ikincil yapı tahmininde, bazı çalışmalarda çoklu dizi hizalama yöntemleri kullanıldığı gibi, son zamanlarda tüm protein dizisini eğitim girdisi olarak kullanan çalışmalar da bulunmaktadır. En sık kullanılan çoklu dizi hizalama algoritmaları arasında PSI-BLAST [32] ve HHblits yer almaktadır [33].

Makine öğrenmesi teknikleri dışında, protein ikincil yapı tahmininde başarı oranını arttırmak için birden fazla başarılı metodun bir arada kullanıldığı hibrid yöntemler geliştirilmiştir [34], [35].

Karmaşık problemleri çözmedeki başarısı ve büyük veri setleriyle çalışma imkânı ile derin öğrenme yöntemleri pek çok alanda olduğu gibi biyoinformatik alanında yapılan çalışmalarda da sıklıkla tercih edilmiştir [36], [37]. Bu alandaki çalışmalarla birlikte derin öğrenme algoritmalarından Evrişimli Sinir Ağları ve Özyinelemeli Sinir Ağları protein ikincil yapı tahmininde kullanılan önemli metotlar arasında yer almıştır.

90'lı yılların sonlarında Baldi ve arkadaşları tarafından protein ikincil yapılarının tahmininde PSI-BLAST hizalaması ve çift yönlü (bidirectional) özyinelemeli sinir ağı kullanılarak geliştirilen SSPro, üç sınıflı tahminlemede %76 başarı elde etmiştir [38]. Pollastri ve McLysaght SSPro'yu geliştirilerek Q_3 başarı oranı %79 olan Porter adını verdikleri bir web sunucusu oluşturmuşlardır [39]. Sonraki yıllarda Porter'ın çeşitli versiyonları da geliştirilmiştir. Porter 4.0 adı verilen çalışmada Q_3 başarı oranı %82,2 olarak belirtilmiştir [40].

Heffarnan ve arkadaşları, SPIDER3 adını verdikleri yöntemde, çift yönlü özyinelemeli sinir ağı ve LSTM hücreleri kullanılarak %84,48 Q_3 başarı elde etmişlerdir [41].

Wang ve arkadaşları Şartlı Rastgele Alanlar (Conditional Random Fields-CRF) ve sinir ağları kullanarak geliştirdikleri DeepCNF yöntemi ile üç sınıflı ve sekiz sınıflı tahminleme çalışmaları yapmışlardır [42]. Bu çalışmada, beş farklı test seti ile yapılan değerlendirme sonuçlarına göre Q_3 başarılarının %82,3 ile %85,4 arasında yer aldığı belirtilmiştir.

Wang ve arkadaşları, SSREDNs (Secondary Structure Recurrent Encoder–Decoder Networks) adını verdikleri yöntem ile eğittikleri ağda, CB513 ve CullPDB veri setlerini test ederek Q_3 için %82,9 ve %84,2, Q_8 için 68.20% ve 73.1% başarıya ulaşmışlardır [43].

Son yıllarda, CNN ve RNN gibi derin öğrenme yöntemleri ile başarılı tahmin sonuçları elde edilmesi ile birlikte bu yöntemlerin bir arada kullanıldığı çalışmalarda tahmin başarısının arttırılması hedeflenmiştir [44].

Li ve Yu, DCRNN (Deep convolutional recurrent neural network) adını verdikleri çok ölçekli (multiscale) CNN ve çift yönlü geçitli tekrarlayan birim katmanları kullandıkları çalışmada farklı veri setlerinde yaptıkları testlerde CB513 veri seti için %69,7, CASP10 veri seti için %76,9 ve CASP11 veri seti için %73,1 Q_8 başarısı elde etmişlerdir [45]. Ayrıca bu çalışmanın, literatürde PİYT için GRU kullanıldığı bilinen ilk çalışma olduğunu belirtmişlerdir.

Guo ve arkadaşları tarafından yapılan çalışmada, çift yönlü uzun-kısa süreli bellek ve asimetrik CNN kullanılarak sekiz sınıflı protein ikincil yapı tahmin çalışması yapılmıştır [46]. Çalışma sonuçlarına göre, DeepACLSTM adı verilen bu algoritma ile CASP10 veri setinde %75, CASP11 veri setinde %73 Q_8 başarısına ulaşmışlardır.

Wang ve arkadaşları, 13 pencere boyutlu PSSM matrisini eğitim girdisi olarak kullanarak CNN ve LSTM katmanlarından oluşan ağ ile eğitmişler ve %80,18 Q_3 başarısı elde etmişlerdir [47].

BÖLÜM 3. DERİN ÖĞRENME

Yapay zekâ, zeki canlılar tarafından kolaylıkla gerçekleştirilebilen süreçlerin, makineler veya bilgisayarlar tarafından bu canlılara benzer şekilde gerçekleştirilme yeteneği olarak tanımlanmaktadır. Genellikle insan zekasını taklit eden sistemler olarak düşünülse de sadece bununla sınırlandırılmamaktadır. 1950'li yıllardan itibaren literatürde karşılaşılan bu yöntemler, ilk zamanlarda çoğunlukla, makinelerin problem çözme ve tahmin etme süreçlerinin insanlar tarafından matematiksel ve mantıksal olarak programlanmasına dayanmaktaydı.

Yapay zekânın alt kümesi olan ve 1980'lerde yükselişe geçen makine öğrenmesi kavramı ise, Arthur Samuel tarafından makinelerin özellikle programlanmadığı sonuçları öğrenebilme yeteneği olarak tanımlanmıştır [48]. Bilgisayar sistemlerinin, donanımsal gelişimi ile birlikte bu yöntemlerin gerçekleştirilebilmesi daha mümkün hale gelmiştir. Makine öğrenmesi yöntemlerinin başında, yapay sinir ağları (YSA), destek vektör makineleri (SVM- Support Vector Machine) ve genetik algoritmalar gelmektedir.

Yapay sinir ağları, insanların öğrenmesi esnasında beyinde gerçekleşen süreçlerde sinir hücrelerinin ve aralarındaki iletişimin modellendiği bir yapıya sahiptir. YSA'da nöronlar arasındaki bağlantı ağırlıklarla hesaplanır. Makinenin öğrenmesi istenen girdi ve çıktılar arasındaki bu bağlantı ileri veya geri beslemeli olarak hesaplanabilir.

Örneklerden öğrenen makine öğrenmesi modellerinde, verinin yani örnek miktarının artması öğrenmeyi olumlu yönde etkilemektedir. Ancak zaman içerisinde tek katmanlı bir öğrenme sistemi sağlayan YSA gibi yöntemlerin çok miktarda veriyi öğrenmesi için güçlü sistemler ve zaman gerekmiştir. Bu sebeple, çok katmanla çalışan derin öğrenme yaklaşımı günümüzde sıklıkla tercih edilmeye başlanmıştır. Derin öğrenme, yapay sinir ağlarına dayanan bir makine öğrenmesi yaklaşımıdır.

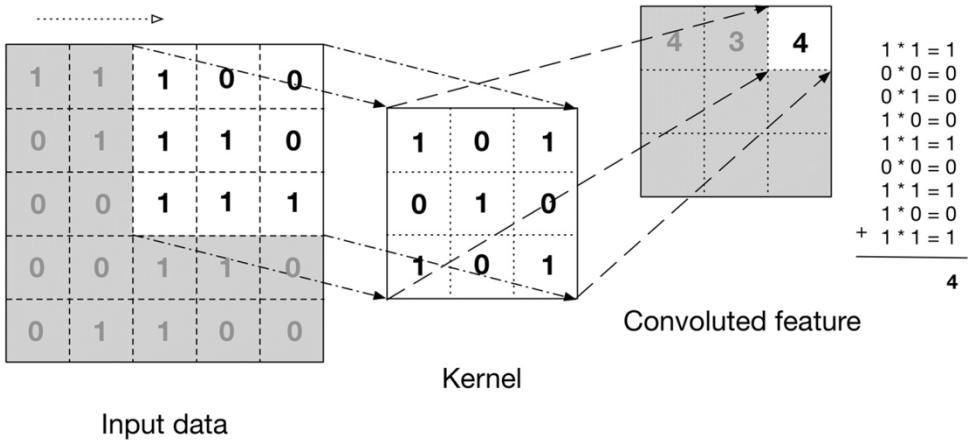
Tezin bu bölümünde, bu çalışmada kullanılan derin öğrenme yöntemleri ile ilgili teknik bilgilere yer verilecektir.

3.1. Evrişimli Sinir Ağları

Evrişimli sinir ağları (CNN), çok boyutlu dizi şeklindeki verilerden öğrenim yapmak için kullanılan bir derin öğrenme modelidir. Genellikle resim gibi iki boyutlu verilerde yer alan bilginin, resimdeki figürler veya cisimler gibi, öğrenilmesinde sıklıkla CNN yapıları kullanılır. Bu yapıda, örnekler girdi olarak ağa iletildikten sonra, gizli katmanlarda çıktının diğer bir deyişle istenen bilginin öğrenimi gerçekleşir. Verinin boyutuyla ilişkili olarak genellikle, dizi ve sekans verilerinin öğreniminde tek boyut, resim ve ses dalgalarında iki boyut, video için üç boyutlu evrişimli sinir ağları kullanılır. Evrişimli sinir ağlarında kullanılan katmanlar temel olarak, evrişim, örnekleme (pooling) ve tam bağlı katmanlar olarak karşımıza çıkmaktadır.

3.1.1. Evrişim katmanı

Bu katman, konvolüsyon olarak da adlandırılan evrişim işlemiyle özellik çıkarımının yapıldığı CNN modellerinin temeli sayılan ilk katmandır ve arka arkaya birden fazla eklenerek tekrarlayabilir. İşlenmemiş veriler gibi önceki katmanın çıktısı olan özellik haritaları da bu katmanın girdisi olabilir. Konvolüsyon işleminde, kernel olarak adlandırılan bir filtre aracılığıyla girdide öne çıkan özelliklerin çıkarımı yapılır. *NxN* boyutlu bu filtre ile girdi matrisinde soldan sağa ve yukarıdan aşağı doğru adım adım girdi verisi ile filtredeki veri arasında çarpma işlemi gerçekleştirilip tüm değerler toplanarak bir değer elde edilir. Girdi verisi filtre ile bu şekilde gezildikten sonra özelliklerin belirlenmiş olduğu bir çıktı matrisi oluşturulmuş olur. Özellik haritası olarak da adlandırılan bu matris, bir sonraki katmanda girdi olarak kullanılır. Şekil 3.1.'de konvolüsyon işleminin uygulanışı gösterilmektedir.

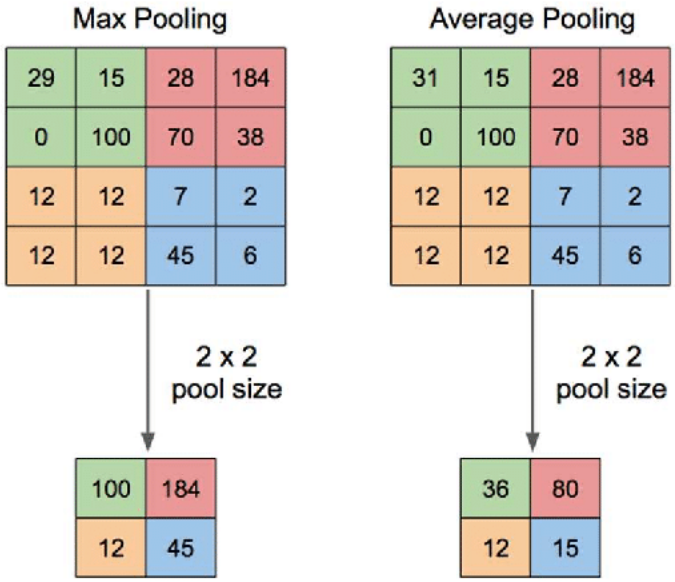


Şekil 3.1. Konvolüsyon işleminin uygulanışı [49]

3.1.2. Örnekleme

Evrişimli sinir ağlarında sıklıkla yer alan diğer bir katman olan pooling katmanı, genellikle iki evrişim katmanı arasında bulunur. Hesaplanan parametreleri azalmak amacıyla bir önceki katmandan gelen özellik haritasında *nxn* boyutlu bir filtre ile işlem yapılarak önemli özelliklerin belirlenmesi sağlanır. Genellikle maximum pooling ve ortalama pooling olarak adlandırılan iki türü kullanılmaktadır. Maksimum pooling ile filtreye karşılık gelen özellik haritası parçasındaki en büyük değer alınırken, ortalama ile ortalama değer hesaplanır. Bu işlem sonucunda, matris boyutu küçülürken önemli özellikler öne çıkmış olur.

Şekil 3.2.'de girdi matrisi üzerinde *2x2* boyutlu filtrenin 2 adım aralığı ile gezilerek öznitelik matrislerinin çıkarılması işlemi iki farklı örnekleme türü kullanılarak gösterilmektedir. İlk adımda, maksimum örnekleme ile filtreye karşılık gelen matris bölümündeki en büyük değer olan 100 öznitelik olarak seçilirken, ortalama örneklemede 31,15, 0 ve 100 değerlerinin ortalaması alınmaktadır.



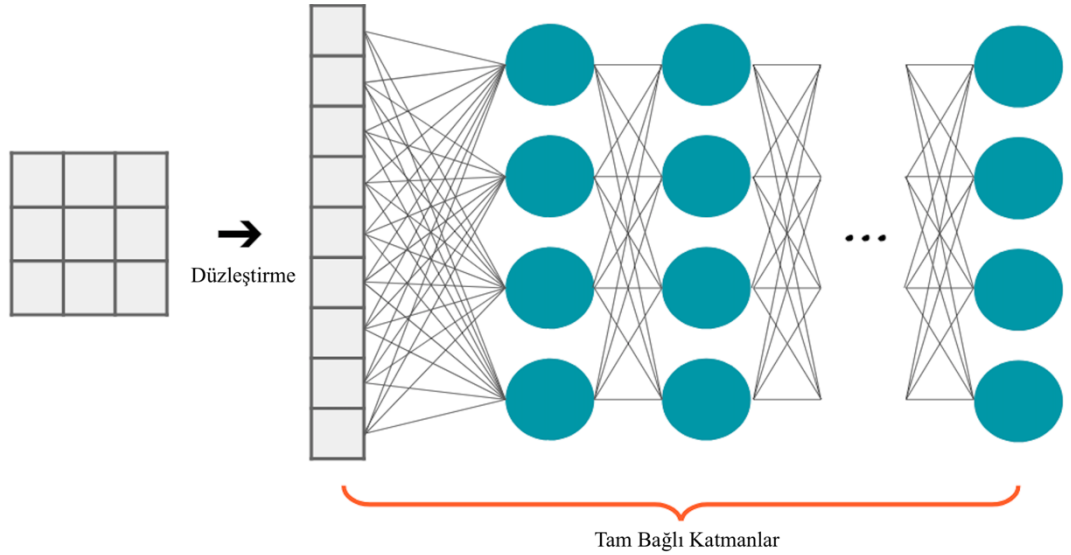
Şekil 3.2. Maksimum ve ortalama örnekleme işleminin uygulanışı [50]

3.1.3. Düzleştirme

Önceki başlıklarda belirtilen katmanlarda yapılan işlemler sonucu elde edilen özellik matrisi, düzleştirme işlemi ile tek boyutlu bir dizi haline getirilerek, sınıflandırma işlemini gerçekleştirecek son katmanlar için hazırlanır. Düzleştirme işlemi tam bağlı katmanlarından önce yer alır.

3.1.4. Tam bağlı katman

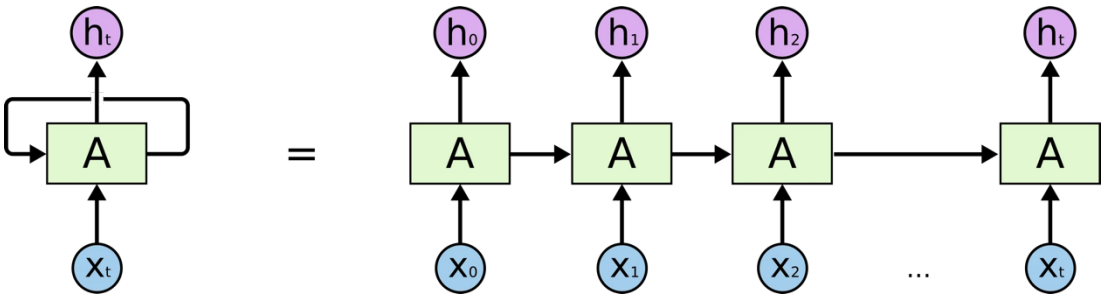
Tam bağlı katman (Fully Connected Layer), CNN mimarilerinde yer alan son katmanları temsil eder. Bu katmanlarda, klasik yapay sinir ağlarında olduğu gibi tüm nöronlar arasında bağlantı vardır ve çıkış katmanından önceki optimizasyon işlemleri gerçekleştirilir. Tam bağlı katmanda, girdi olarak bir vektör gönderilmesi gerektiğinden bu katmandan önce düzleştirme işlemine gereksinim duyulur. Şekil 3.3.'te düzleştirme işlemi ve tam bağlı katmanların temsili gösterimi yer almaktadır.



Şekil 3.3. Düzleştirme ve tam bağlı katmanlar [51]

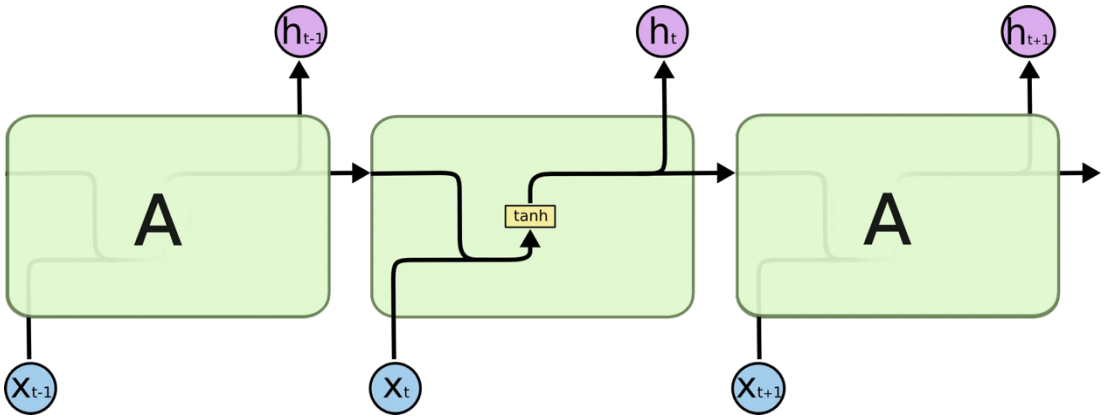
3.2. Tekrarlayan Sinir Ağları

Tekrarlayan sinir ağları (RNN), geçici bir hafızaya sahip ileri beslemeli sinir ağlarıdır. Yapay sinir ağları her bilgiyi ayrı ayrı değerlendirirken, RNN'ler önceki bilgiyi saklayabildiklerinden yeni bilgiyi işlerken önceki bilgilerden faydalanabilirler. Sıralı dizilerin öğrenilmesinde kullanılan RNN, bellek olarak düşünülebilecek gizli durumlara (hidden states) sahiptir. Bir sonraki zaman adımındaki diziyi tahmin etmek için, mevcut bir zaman adımındaki bir giriş ve çıkış dizisini eşleyebilen geri yayılım kullanarak dizi öğrenimini gerçekleştirirler. Yapılarında bulunan döngüler ile sıralı veriler arasında ilişki kurma ve anlamlandırma yapabilirler. Böylelikle, eski ve yeni girdi arasında bilgi akışı sağlanmış olur.



Şekil 3.4. Tekrarlayan sinir ağları [52]

Şekilde gösterilen RNN yapısında *t* zamanındaki durum için, x_t girdiyi, h_t çıktıyı ve A ağın bir kısmını temsil ederken; döngü, öğrenilen bilginin farklı bir *t* zamanında kullanılmasına olanak verir. (Şekil 3.4.)



Şekil 3.5. RNN ağ yapısı [52]

Şekil 3.5.'te döngünün A ile temsil edilen ağ yapısı gösterilmektedir. RNN'lerde aktivasyon fonksiyonu olarak genellikle tanh fonksiyonu kullanılır.

3.3. Uzun-Kısa Süreli Bellek Ağları

Yapay sinir ağlarında sıklıkla kullanılan geri yayılım hesaplamaları, RNN yapısında eski bilgilerin aşırı birikmesi ve t zamanının artması veya dizinin uzaması ile zamanla kaybolması durumunu ortaya çıkarabilmektedir. Uzun-Kısa Süreli Bellek Ağları (LSTM), kaybolan gradyan (vanishing gradient) olarak adlandırılan bu problemin çözümü olarak önerilen bir RNN türüdür. İlk olarak 1997 yılında Hochreiter & Schmidhuber tarafından tanıtılan LSTM, kısa süreli öğrenimleri hafıza bloklarında tutarak uzun süreli kullanımını sağlamaktadır [53].

LSTM'ler, girdi, çıktı ve gizli katmanlardan oluşmaktadır ancak RNN'lerden farklı olarak bellekte üç kapı bulunur. Giriş, çıkış ve unutma kapısı olarak adlandırılan bu kapılar, bellekten bilgiyi okuma, yazma ve güncelleme işlemlerini yürütmektedirler.

3.4. Geçitli Yinelenen Birim

Geçitli Yinelenen Birim (GRU), Cho ve arkadaşları tarafından kaybolan gradyan problemine çözüm sağlamak amacıyla 2014 yılında önerilen bir RNN türüdür [54]. LSTM yapısında üç kapı bulunurken, GRU'da iki kapı yer almaktadır. Sıfırlama ve güncelleme olarak adlandırılan kapılar ile, tahmin edilmek istenen çıktılar için gerekli bilgilerin uzun süre bellekte tutulması, gereksiz bilgilerin ise sıfırlanması işlemleri gerçekleştirilir.

$$z_t = \sigma\left(W^{(z)}x_t + U^{(z)}h_{t-1}\right) \tag{3.1}$$

Denklemde, *t* zamanındaki güncelleme adımının formülü gösterilmektedir. *t* zamanındaki x_t girdisi kendi ağırlığı olan $W^{(z)}$ ile çarpılır. İşlemin yapıldığı zamandan önceki *t-1* zamanında elde edilen bilgi aynı şekilde kendi ağırlığı ile çarpılarak, elde edilen değerlere toplama işlemi uygulanır. Bu sonucun, sigmoid aktivasyon fonksiyonu uygulanarak 0-1 aralığına dönüştürülmesi ile z_t değeri elde edilir.

$$r_t = \sigma\left(W^{(r)}x_t + U^{(r)}h_{t-1}\right) \tag{3.2}$$

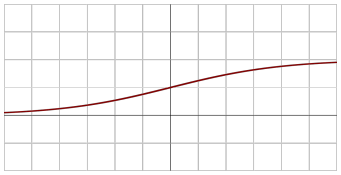
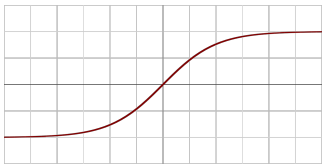
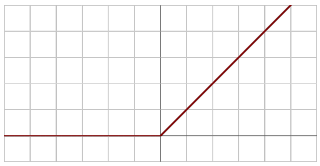
Yukarıdaki denklem, GRU'larda yer alan unutma işleminin formülü gösterilmektedir. Unutma kapısında, önceki girdilerden öğrenilen bilgilerin ne kadarının unutulacağı belirlenmiş olur.

3.5. Aktivasyon Fonksiyonları

Transfer fonksiyonu olarak da adlandırılan aktivasyon fonksiyonları, yapay sinir ağları ve derin öğrenme modellerinde karmaşık verilerin öğrenilmesini sağlayan fonksiyonlardır. Girdi verilerinden üretilecek çıktının hesaplanması bu fonksiyonlar ile gerçekleştirilir. Aktivasyon fonksiyonları, verilerin belirli değer aralığına dönüştürülmesi işlemini yürütür. Derin öğrenme çalışmalarında sıklıkla kullanılan aktivasyon fonksiyonları; ReLu, tanh, sigmoid olarak sıralanabilir.

Tablo 3.1.'de, bu çalışmada yer alan aktivasyon fonksiyonlarının grafikleri ve denklemleri gösterilmektedir.

Tablo 3.1. Aktivasyon fonksiyonları

Fonksiyon Adı	Grafik	Denklem
Sigmoid		$\sigma(\mathrm{x}) = \frac{1}{1 + e^{-\mathrm{x}}}$
Hiperbolik Tanjant (tanh)		$\tanh(\mathrm{x}) = \frac{e^{\mathrm{x}} - e^{-\mathrm{x}}}{e^{\mathrm{x}} + e^{-\mathrm{x}}}$
Rektifiye Doğrusal Birimi (ReLu)		$\begin{cases} 0 & \mathrm{x} \leq 0 \text{ için} \\ \mathrm{x} & \mathrm{x} > 0 \text{ için} \end{cases}$

RNN, LSTM ve GRU modellerinde en çok kullanılan aktivasyon fonksiyonu hiperbolik tanjant (tanh) fonksiyonudur. Bu fonksiyon, [-1,1] arasında değer üretir ve doğrusal değildir. Uç değerler için fonksiyonun türevi sıfıra yakınsar.

Özellikle evrişimli katmanlarda sıklıkla tercih edilen Rektifiye Doğrusal Birimi (ReLu-Rectified Lineer Unit),), negatif değerler için 0, pozitif değerler içinse aynı değeri üreten doğrusal olmayan bir aktivasyon fonksiyondur. [0, +∞) arasında değer üretilir. Diğer aktivasyon fonksiyonları ile karşılaştırıldığında, negatif değerlerin sıfırlanması daha hızlı çalışmasını sağlamaktadır.

Sınıflandırma algoritmalarında, girdinin belirlenen sınıfa ait olma olasılığını hesaplamayı sağlayan softmax aktivasyon fonksiyonu genellikle çıktı katmanında tercih edilir. Softmax kullanılırken, çıktı katmanında problemde yer alan sınıf kadar nöron sayısı olması gerekmektedir.

Aşağıdaki denklemde $i = 1, \dots, K \ için$ softmax fonksiyonu gösterilmektedir.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{3.3}$$

3.6. Başarım İyileştirme Yöntemleri

Derin öğrenme modellerinde sıklıkla karşılaşılan veri setini aşırı öğrenme veya ezberleme (overfitting) durumlarını engellemek ve model başarısını arttırmak için çeşitli regülasyon yöntemleri tercih edilmektedir. Bunlardan en sık kullanılan iki yöntem seyreltme (dropout) ve L2 regülasyonudur.

3.6.1. Seyreltme

Eğitim sırasında, düğümler arasındaki bağlantıların belirlenen eşik değeri doğrultusunda veya rastgele koparılması seyreltme (dropout) işlemi olarak tanımlanmaktadır. Srivastava ve arkadaşları tarafından 2014 yılında önerilmiştir ve aşırı öğrenme problemlerine çözüm olarak kullanılabildiği gibi, model başarısını arttırmak için de kullanılabileceği belirtilmiştir [55].

3.6.2. L2 düzenleme

L2, cezalandırma yöntemi uygulayarak ağırlıkları çok yüksek olmamasını sağlayan bir düzenleme yöntemidir. Model ağırlıklarını düzenleyerek aşırı ezberleme durumunun engellenmesine ve model başarısını arttırılmasına yardımcı olur.

3.7. ADAM Optimizasyonu

Derin öğrenme ve YSA modellerinde ağırlık değerlerinin güncellenmesi ve hatanın en aza indirgenmesi için çeşitli optimizasyon algoritmaları kullanılmaktadır. 2014 yılında Kingma ve arkadaşları tarafından önerilen Adam, birinci dereceden gradyan tabanlı bir eniyileme algoritmasıdır [56].

BÖLÜM 4. MATERYAL VE YÖNTEM

4.1. Veri Seti

Bu tez çalışmasında, Cuff ve Barton [57] tarafından oluşturulan, 513 protein ve 84.119 amino asit içeren açık bir veri seti olan CB513 kullanılmıştır. PSI-BLAST ve HHBlits hizalama yöntemleri kullanarak elde edilen PSSM öznitelik vektörleri ve DSPRED yöntemi ile elde edilen yapısal profil matrislerini kullanarak düzenlenen CB513 veri seti, Aydın ve arkadaşlarının[58] çalışmasından elde edilmiştir.

Veri setinde, her bir amino asit 539 öznitelikle temsil edilmektedir. Bu öznitelikler, CB513 veri setinde yer alan her bir proteinin amino asit diziliminin belirtilen yöntemlerle hizalanması sonucunda 20xN boyutlarında iki adet PSSM matrisi ve proteinin ikincil yapısını temsil eden 3xN boyutunda üç adet yapısal profil matrisi ile elde edilen 49 özniteliğin, hedef amino asitlerin çevresindeki amino asitlerle etkileşimini ölçmek amacıyla 11 birim uzunluğundaki pencerelerin kullanılmasıyla elde edildiği belirtilmiştir [59]. Doğada bilinen 20 çeşit amino asit bulunduğundan proteinler, 20xN olarak temsil edilir. Proteinlerin içerdikleri amino asit sayıları farklı olabildiğinden, N hedef proteinde yer alan amino asit sayısını belirtir. Benzer olarak, üç sınıflı protein ikincil yapı tahmini çalışmalarında proteinlerin ikincil yapısı, 3xN boyutundaki matrislerle temsil edilmektedir.

Kullanılan veri seti, 513 proteinin, her bir katın daha sonra test için kullanıldığı yedi çapraz doğrulama setinden oluşmaktadır. K-katlamalı çapraz doğrulama, modelin başarımını test etmek için kullanılmaktadır. Bu, her modelin 7 kez farklı veri kümesi kullanılarak eğitileceği ve test edileceği anlamına gelmektedir.

4.2. Geliştirme Ortamı

4.2.1. Google Colaboratory

Google Colaboratory, özellikle makine öğrenmesi, veri analizi, derin öğrenme gibi güçlü donanımsal ekipman gerektiren uygulamalar için ücretsiz GPU erişimi sunan, bu uygulamaların Python dili ile bir internet tarayıcısı üzerinden geliştirilmesine olanak tanıyan bulut tabanlı bir platformdur [60]. Bu tez çalışmasında gerçekleştirilen uygulamaların tümü bu platform üzerinden GPU kullanılarak yürütülmüştür. GPU üzerinde yürütme işlemleri için Colaboratory kullanıcılara 12 saat aralıksız kullanım hakkı tanımlamaktadır, bu sürenin sonunda geçici kısıtlama uygulanmakta ve kısıtlama sonunda tekrar GPU kullanımına izin vermektedir. Ancak GPU kullanımı, çalışma süresini epey hızlandırdığından bahsedilen süre çalışmaların yürütülmesi için yeterli olmuştur.

4.2.2. Kullanılan kütüphaneler

Bu tez çalışmasında, üç sınıflı protein ikincil yapı tahmini yapmak için beş farklı derin öğrenme ağı Python programlama dili kullanılarak Colaboratory platformu üzerinde geliştirilmiştir. Verilerin aktarılması, önişleme, derin öğrenme ağının tasarlanması, eğitim, test ve tüm değerlendirme aşamaları için çeşitli kütüphaneler kullanılmıştır. Uygulamaların gerçekleştirilmesi sırasında başlıca; NumPy, Pandas, Scikit-Learn, Seaborn, Matplotlib ve Keras kütüphaneleri kullanılmıştır.

NumPy, dizi ve matrisler ile matematiksel işlemler yapılmasına imkân veren açık kaynak kodlu bir kütüphanedir. Pandas, verilerin farklı dosya türlerinden okunması, bu verilerin analiz edilmesi ve tekrar dosyaya yazılması gibi birçok işlemin kolaylıkla gerçekleştirilebildiği bir veri analizi aracıdır. Çalışmada kullanılacak verilerin yüklenme ve düzenlenme işlemleri için NumPy ve Pandas kütüphaneleri kullanılmıştır.

Scikit-Learn, farklı kütüphanelerle entegre çalışabilen bir makine öğrenmesi kütüphanesidir. Bu çalışmada, model başarılarının ölçülmesi için gereken hesaplamalar bu kütüphane kullanılarak elde edilmiştir. Seaborn ve Matplotlib kütüphanelerine, veri görselleştirme ve grafik oluşturmak için başvurulmuştur.

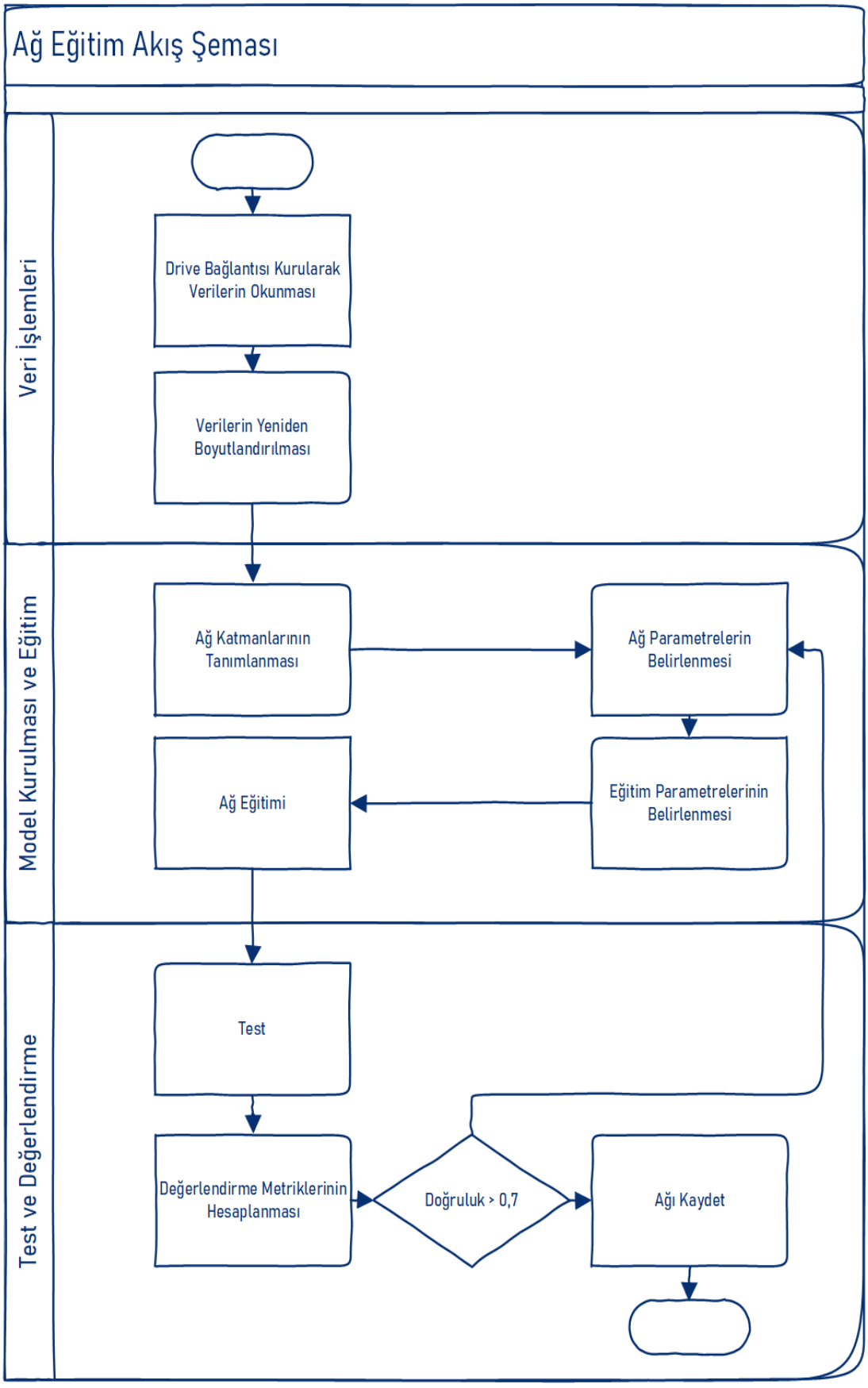
Keras, Tensorflow veya Theano kütüphaneleri üzerinden çalışabilen, Python programlama dili ile yazılmış açık kaynak kodlu bir derin öğrenme kitaplığıdır. Bu tez çalışmasında bahsedilen tüm derin öğrenme ağları, Keras kütüphanesi kullanılarak oluşturulmuştur.

4.3. Geliştirilen Derin Öğrenme Modelleri

Bu tez çalışmasında, ikinci bölümde bahsedilen proteinin birincil yapısından ikincil yapının tahmini için CNN, RNN, LSTM ve GRU olmak üzere dört farklı derin öğrenme ağı oluşturulmuştur. Modellerin oluşturulması ve ağ yapıları bu başlık altında paylaşılmıştır.

Modellerin eğitiminde kullanmak için öncelikle verilerin bulunduğu Google Drive ile Colaboratory arasında bağlantı kurulmuş ve veriler programa aktarılmıştır. Kullanılan veri seti için 7 katlı çapraz doğrulama uygulandığından, eğitim için kullanılacak veriler, her çapraz doğrulama seti için, 11 pencere boyutunu ve her amino asit için belirlenen 49 özniteliği temsil etmek için N veri setinin satır sayısı olmak üzere Nx49x11 olarak yeniden boyutlandırılmıştır. Her kat için ayrılan eğitim verisinin %10'u, eğitim esnasında kullanılmak üzere doğrulama seti olarak ayırılmıştır.

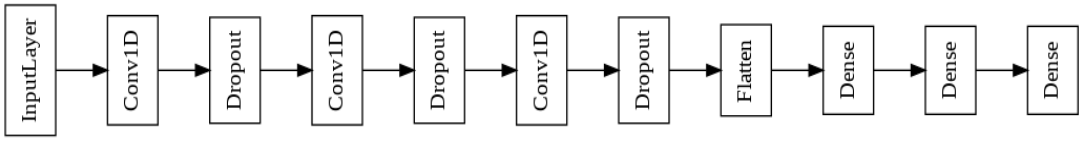
Veriler eğitim için hazırlandıktan sonra, ağ katmanları ve parametreleri tanımlanmış ve model eğitilmiştir. Modelin eğitiminden sonra eğitim ve doğrulama testlerinin başarısı ve kayıp fonksiyonları incelenmiştir. Test verisinin çıktıları, eğitilen ağ ile tahmin edilip, olması gereken ikincil yapı sınıfları ile tahmin edilenler sınıfları karşılaştırmak için değerlendirme metrikleri hesaplanmış ve grafikler hazırlanmıştır. Tüm modeller için gerçekleştirilen eğitim süreci aynı işlemleri gerektirmektedir. Modellerin ağ eğitimini gösteren akış şeması Şekil 4.1.'de gösterilmiştir.



Şekil 4.1. Ağ eğitim akış şeması

4.3.1. CNN modeli

Protein ikincil yapı tahmini çalışması yapmak üzere geliştirilen Evrişimli Sinir Ağları modeli, üç evrişim katmanı arasında seyrelme katmanları ve ardından düzleştirme katmanı ile tam bağlı katmanlardan oluşmaktadır. Ağ katmanları Şekil 4.2.'de yer almaktadır.



Şekil 4.2. Geliştirilen CNN modelinin ağ katmanları

Bu çalışmada kullanılan CNN ağının evrişimli katmanlarında sırasıyla, 128, 64, 32 filtre ile 5, 3, 3 kernel ve ReLu aktivasyon fonksiyonu kullanılmıştır. Evrişim katmanlarında kernel regülasyonu olarak 0,001 değerinde L2 regülasyonu eklenmiştir. Evrişimli katmanlar arasında yer alan seyretme katmanları için 0,20 değeri belirlenmiştir. Proteinin ikincil yapısını temsil eden üç sınıfın son katmanda tahmin edilmesi için çıkış boyutu üç olarak belirlenmiş ve softmax fonksiyonu kullanılmıştır.

Tablo 4.1. Geliştirilen CNN modelinin ağ yapısı ve parametreleri

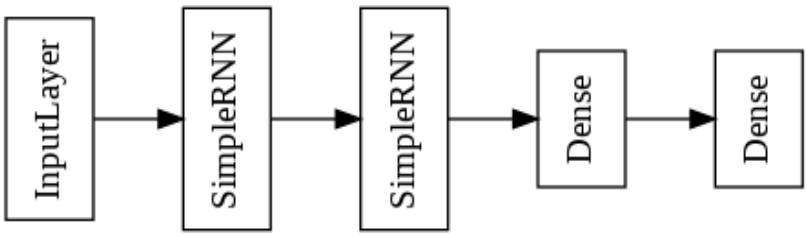
Katman Türü	Çıktı Boyutu	Parametre	Aktivasyon
Evrişim Katmanı (Conv1D)	(None, 45, 128)	7168	relu
Seyreltme (Dropout)	(None, 45, 128)	0	
Evrişim Katmanı (Conv1D)	(None, 43, 64)	24640	relu
Seyreltme (Dropout)	(None, 43, 64)	0	
Evrişim Katmanı (Conv1D)	(None, 41, 32)	6176	relu
Seyreltme (Dropout)	(None, 41, 32)	0	
Düzleştirme (Flatten)	(None, 1312)	0	
Dense	(None,64)	84032	relu
Dense	(None,32)	2080	relu
Dense	(None,3)	99	softmax

Toplam parametre sayısı : 124195
Eğitebilir parametre sayısı. : 124195
Eğitilemez parametre sayısı : 0

Tablo 4.1.'de belirtilen ağ yapısı ve parametreleri belirlendikten sonra, toplamda 124195 parametre eğitilmiş ve modelin eğitimi 20 eğitim turu (epoch) sonunda tamamlanmıştır. Eğitimde kullanılan öğrenme katsayısı 0,0001, optimizasyon algoritması Adam ve yığın boyutu 64 olarak belirlenmiştir.

4.3.2. RNN modeli

İkincil yapı sınıflarının tahmini için oluşturulan ikinci model, tekrarlayan sinir ağlarını kullanılarak geliştirilmiştir. Girdi katmanı, iki RNN katmanı ve sonda yer alan çıkış katmanı olmak üzere iki yoğun katmandan oluşmaktadır. Ağ katmanlarının temsili gösterimi Şekil 4.3.'te gösterilmiştir.



Şekil 4.3. Geliştirilen RNN modelinin ağ katmanları

Ağ yapısında bulunan, iki adet RNN katmanı 64 nöron ile yoğun katman 32 nöron içermekte ve relu aktivasyon fonksiyonu kullanmaktadır. CNN modelinde belirtildiği gibi, çıkış katmanın boyutu üç ve aktivasyon fonksiyonu softmax olarak belirlenmiştir.

RNN katmanlarında, sekansı geri döndürme özelliği bulunmaktadır. Bu katmana iletilen girdi verisinin, sonraki katmana iletilmesi anlamına gelmektedir. Sekans döndürme özelliği ilk katmanda kullanılmıştır. İkinci katmandan sonra yoğun katman bulunduğundan ve girdi boyutu sekans boyutuyla uyuşmadığından bu katmanda sekans döndürme kullanılmamıştır. RNN modelinin ağ yapısı Tablo 4.2.'de gösterilmektedir.

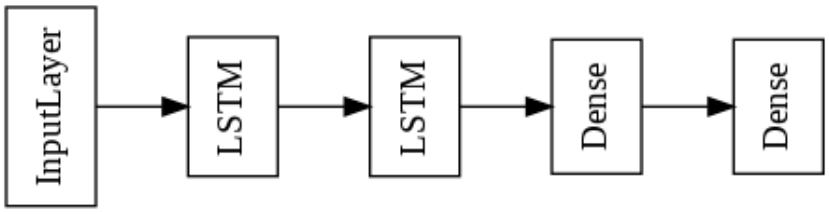
Tablo 4.2. Geliştirilen RNN modelinin ağ yapısı ve parametreleri

Katman Türü	Çıktı Boyutu	Parametre	Aktivasyon	Sekansı Döndür
SimpleRNN	(None, 49, 64)	4864	relu	Açık
SimpleRNN	(None, 49, 64)	8256	relu	Kapalı
Dense	(None,32)	2080	relu	
Dense	(None,3)	99	softmax	
Toplam parametre sayısı : 12495				
Eğitebilir parametre sayısı : 12495				
Eğitilemez parametre sayısı : 0				

Ağ katmanları belirlendikten sonra, eğitim parametreleri belirlenmiştir; öğrenme katsayısı 0,0001, optimizasyon algoritması Adam ve yığın boyutu 64. Toplamda 15299 parametre eğitilmiş ve modelin eğitimi 15 eğitim turu sonunda tamamlanmıştır.

4.3.3. LSTM modeli

Üçüncül yapı sınıflandırma çalışması yapmak için geliştirilen diğer model, iki LSTM katmanı, bir yoğun katman ve çıktı katmanından oluşmaktadır. Ağ katmanları Şekil 4.4.'te gösterilmektedir.



Şekil 4.4. Geliştirilen LSTM modelinin ağ katmanları

RNN modeline benzer olarak, bu modelde iki LSTM katmanı yer alırken, ilk katmanda sekans döndürme özelliği kullanılmıştır. İlk katmanda 128, ikinci katmanda 64 nöron tanımlanmıştır. LSTM modelinin ağ katmanlarında kullanılan parametreler ve aktivasyon fonksiyonları. Tablo 4.3.'te yer almaktadır. Tabloda belirtildiği gibi, LSTM katmanlarındaki aktivasyon fonksiyonu tanh, tam bağlı dense katmanında relu ve çıktı katmanında softmax olarak belirlenmiştir.

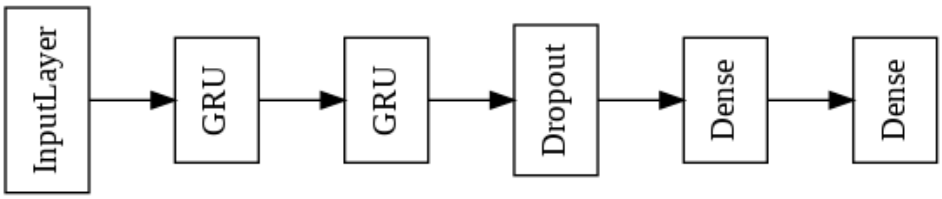
Tablo 4.3. Geliştirilen LSTM modelinin ağ yapısı ve parametreleri

Katman Türü	Çıktı Boyutu	Parametre	Aktivasyon	Sekansı Döndür
LSTM	(None, 49, 128)	71680	tanh	Açık
LSTM	(None, 64)	49408	tanh	Kapalı
Dense	(None,32)	2080	relu	
Dense	(None,3)	99	softmax	
Toplam parametre sayısı : 123267				
Eğitebilir parametre sayısı. : 123267				
Eğitilemez parametre sayısı : 0				

Modelde yer alan 123267 parametre, 20 adımda Adam optimizasyon algoritması, 0,0001 öğrenme katsayısı ve 64 yığın boyutu kullanılarak eğitilmiştir.

4.3.4. GRU modeli

Geliştirilen dördüncü model, Geçitli Yinelemeli Birim modelidir. Bu modelde, Girdi katmanı, iki GRU katmanı ardından seyreltme katmanı ve yoğun katmanlar kullanılmıştır. Şekil 4.5.'te GRU modelinin ağ katmanları gösterilmektedir.



Şekil 4.5. Geliştirilen GRU modelinin ağ katmanları

GRU modeli, ilk iki katmanda sırasıyla 100 ve 50 nörondan oluşan GRU katmanlarını içermektedir. GRU katmanlarından sonra 0,20 oranında seyreltme uygulayan Dropout katmanı yer almaktadır. Son olarak, 50 nörondan oluşan bir yoğun katman ve üç çıkışlı son katman bulunmaktadır. Bu modelde, katmanlarda aktivasyon fonksiyonu olarak tanh kullanılmıştır. Bunun sebebi, GRU katmanlarının varsayılan fonksiyonunun tanh olmasıdır. Aktivasyon fonksiyonu değiştirildiğinde model CPU üzerinde çalışmak zorunda kalır ve eğitim süresi oldukça yavaşlar. Tablo 4.4.'te model katmanları ve parametreleri belirtilmiştir.

Tablo 4.4. Geliştirilen GRU modelinin ağ yapısı ve parametreleri

Katman Türü	Çıktı Boyutu	Parametre	Aktivasyon	Sekansı Döndür
GRU	(None, 49, 100)	33900	tanh	Açık
GRU	(None, 50)	22800	tanh	Kapalı
Seyreltme (Dropout)	(None,50)	0		
Dense	(None,50)	2550	tanh	
Dense	(None,3)	153	softmax	
Toplam parametre sayısı : 59403				
Eğitebilir parametre sayısı : 59403				
Eğitilemez parametre sayısı : 0				

Model, 75153 eğitilebilir parametre içermekte ve eğitim 20 adımda gerçekleştirilmektedir. Adam optimizasyon algoritması ve 0,0001 öğrenme katsayısı kullanılarak eğitilmiştir. Yığın boyutu 64 olarak belirlenmiştir.

4.4. Değerlendirme Metrikleri

Modellerin eğitimleri gerçekleştirildikten sonra, test aşamasına geçilmiştir. Test sonuçlarını etkili biçimde karşılaştırabilmek için sınıflandırma problemlerinde kullanılan metrikler hesaplanmıştır. Metriklerin detayları, bu başlık altında paylaşılmıştır.

Sınıflandırma problemleri değerlendirilirken, modelin tahmin sonucunda ürettiği değerler dört durumda incelenir. Bunlar; doğru tahmin edilen olumlu sınıflar için doğru pozitif (True Positive, TP), yanlış tahmin edilen olumlu sınıflar için yanlış pozitif (False Positive, FP), doğru tahmin edilen olumsuz sınıflar için doğru negatif (True Negative, TN) ve yanlış tahmin edilen olumsuz sınıflar yanlış negatif (False Negative, FN) olarak isimlendirilir. Bu durumlar kullanılarak hesaplanan model değerlendirme ölçütlerinden doğruluk, duyarlılık ve kesinlik bu çalışmada modelleri karşılaştırmak için tercih edilmiştir.

4.4.1. Başarı oranı

Başarı oranı (Accuracy), modelin doğru tahmin ettiği örnek sayısının tüm örnek sayısına oranı ile belirlenen ve sınıflandırma modellerinin genel başarısını değerlendirmek için kullanılan en yaygın ölçüttür.

$$Başarı\ Oranı\ = \frac{TP\ +\ TN}{TP + FP + TN + FN} \tag{4.1}$$

4.4.2. Duyarlılık

Duyarlılık (Recall), doğru tahmin edilen pozitif sınıfların *(TP)*, pozitif tahmin edilmesi gereken tüm durumlara *(TP+FN)* oranını temsil eder.

$$Duyarlılık\ = \frac{TP}{TP + FN} \tag{4.2}$$

4.4.3. Kesinlik

Doğru tahmin edilen olumlu sınıfların (TP), olumlu yani pozitif olarak tahmin edilen tüm durumlara (TP+FP) oranı ile hesaplanan ölçüt kesinlik (precision) olarak tanımlanmaktadır.

$$Kesinlik\ = \frac{TP}{TP + FP} \tag{4.3}$$

4.4.4. F1 skoru

F1 Skoru, duyarlılık ve kesinlik değerlerinin harmonik ortalamasının hesaplanmasıyla elde edilen değerlendirme ölçütüdür. Özellikle sınıflar arası veri dağılımının eşit olmadığı setlerde sıklıkla tercih edilir.

$$F1\ Skoru = \ 2 * \frac{Duyarlılık\ *\ Kesinlik}{Duyarlılık + Kesinlik} \tag{4.4}$$

Bahsedilen performans değerlendirme metrikleri, kesinlik, duyarlılık ve F1 skoru, çok sınıflı tahmin çalışmalarında her sınıf için ayrı ayrı hesaplanmaktadır.

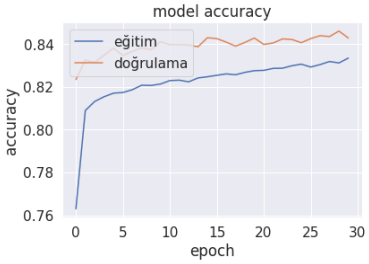
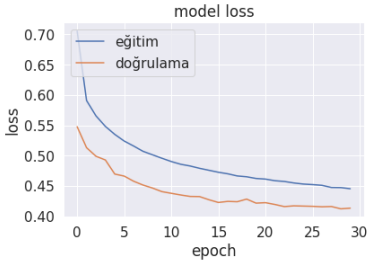
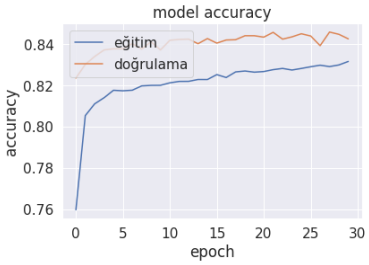
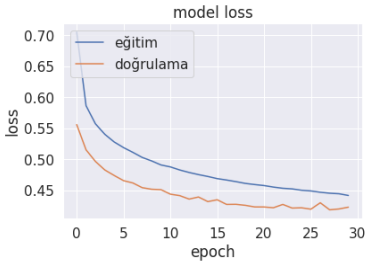
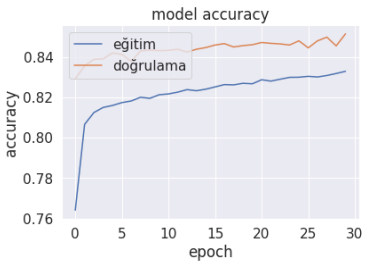
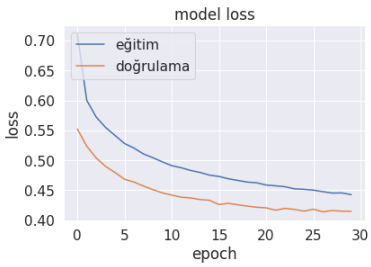
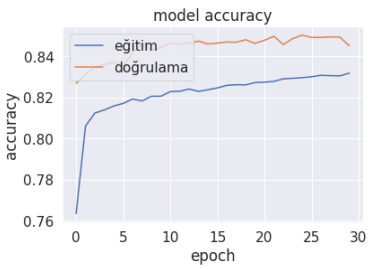
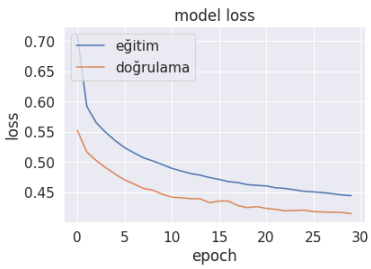
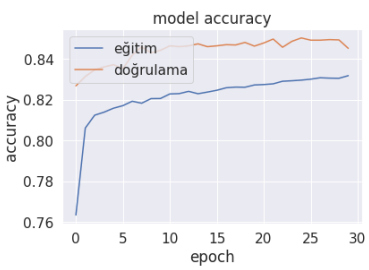
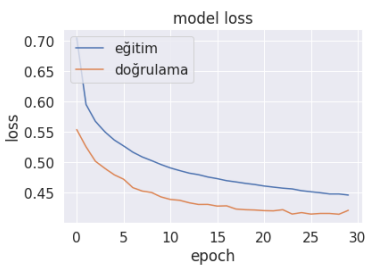
BÖLÜM 5. ARAŞTIRMA BULGULARI

Bu tez çalışmasında, proteinlerin üç sınıflı ikincil yapı tahmini için dört faklı derin öğrenme modeli oluşturulmuş ve modeller aralarında karşılaştırılmıştır. Bölüm 4'te bahsedildiği gibi, modellerin eğitilmesi ve test edilmesi için CB513 veri seti kullanılmıştır. Her model için kesinlik, duyarlılık, F skor ve başarı oranı değerleri hesaplanarak sonuçları değerlendirilmiştir. Modeller, Google Colaboratory hizmeti kullanılarak GPU üzerinde eğitilmiş ve eğitim süreleri karşılaştırılmıştır.

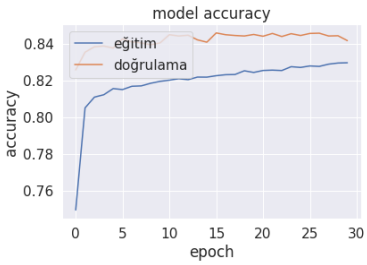
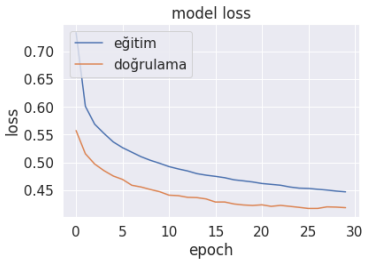
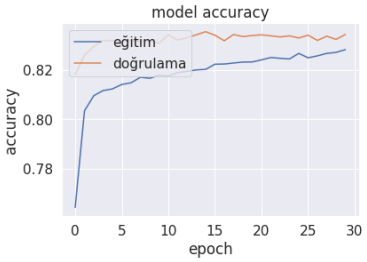
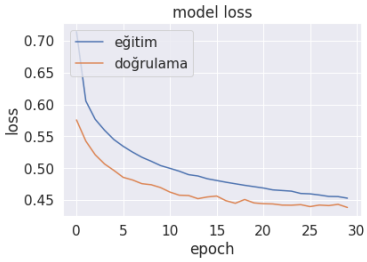
Bu çalışmada oluşturulan dört modelin başarı oranlarının ve değerlendirme metriklerinin ortalamaları hesaplanırken veri setinde uygulanmış olan 7 katlı çapraz doğrulama kümelerinin her biri ile elde edilen sonuçların ağırlıklı ortalaması alınmıştır.

Önceki bölümlerde bahsedildiği gibi, bu çalışmada geliştirilen dört farklı derin öğrenme modeli aynı veri seti kullanılarak eğitilmiştir. Eğitimin tamamlanmasının ardından, eğitim başarı oranının ve kayıp fonksiyonunun belirlenen eğitim adımı süresince gelişimini gösteren grafikler incelenmiş ve eğitimin uyumu kontrol edilmiştir. Her çapraz doğrulama kümesi için CNN, RNN, LSTM ve GRU modellerinin eğitim süresince değişen başarı oranları ve kayıp fonksiyonlarına ait grafikler Tablo 5.1.,5.2.,5.3. ve 5.4.'te yer almaktadır.

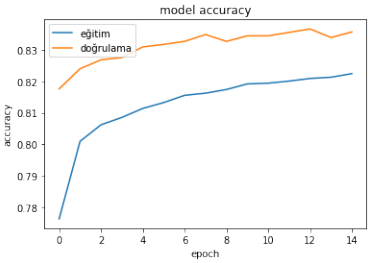
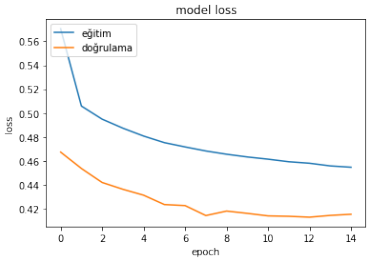
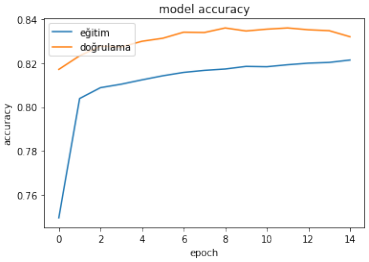
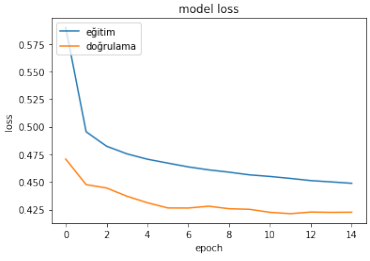
Tablo 5.1. Her bir doğrulama seti için CNN modelinin eğitimi boyunca değişen başarı oranı ve kayıp fonksiyonu değeri grafikleri

	Başarı Oranı	Kayıp Değeri
1		
2		
3		
4		
5		

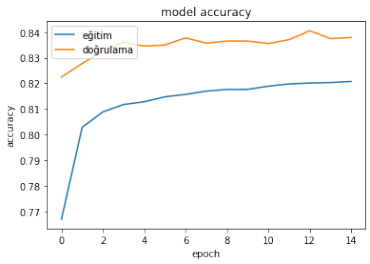
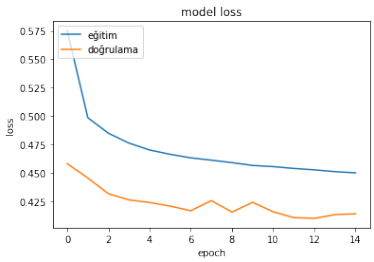
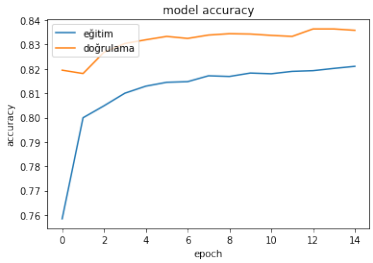
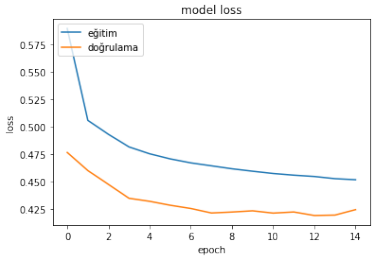
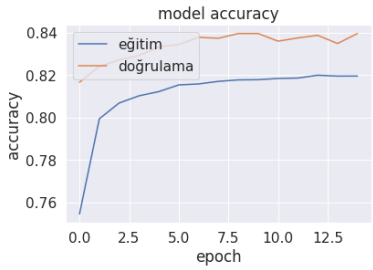
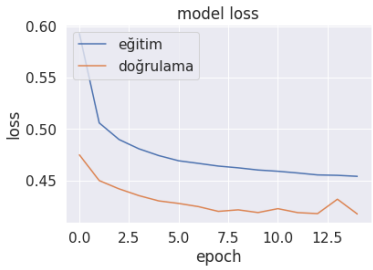
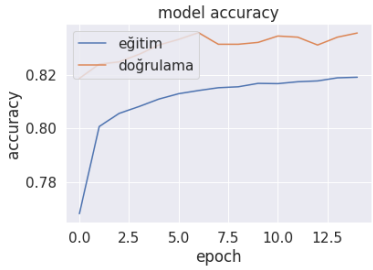
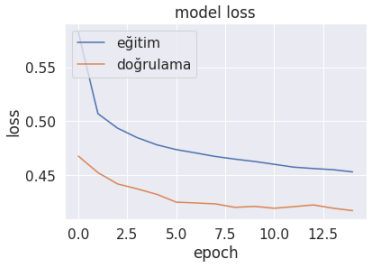
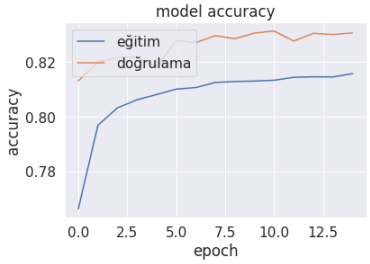
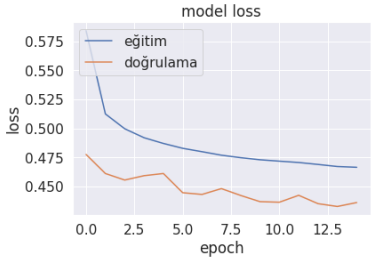
Tablo 5.1. (Devamı)

	Başarı Oranı	Kayıp Değeri
6		
7		

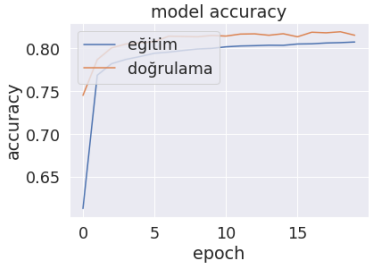
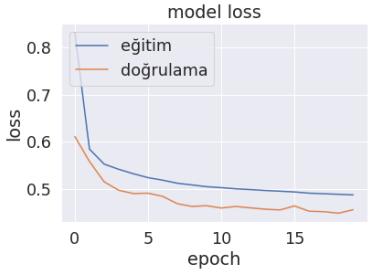
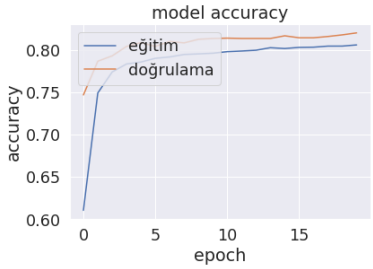
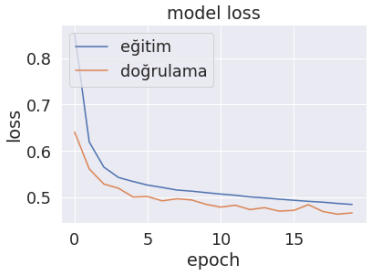
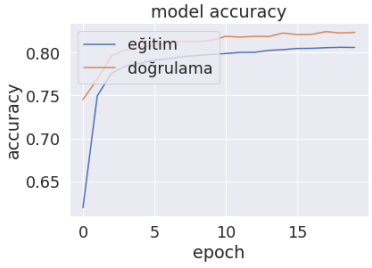
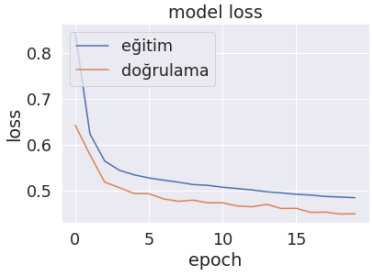
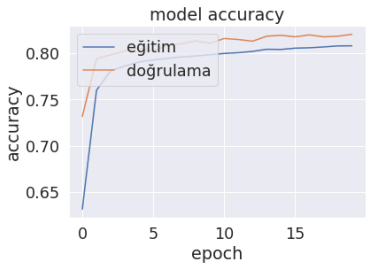
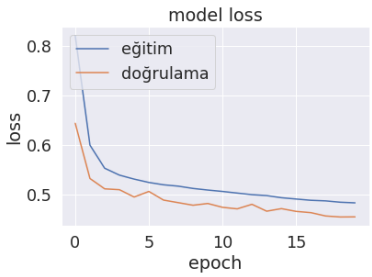
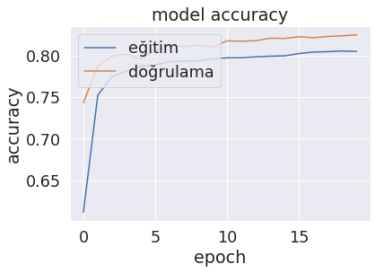
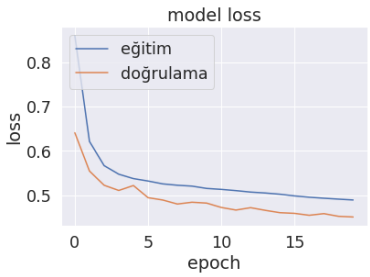
Tablo 5.2. Her bir doğrulama seti için RNN modelinin eğitimi boyunca değişen başarı oranı ve kayıp fonksiyonu değeri grafikleri

	Başarı Oranı	Kayıp Değeri
1		
2		

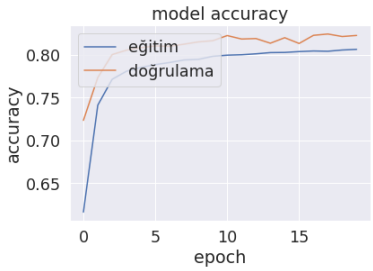
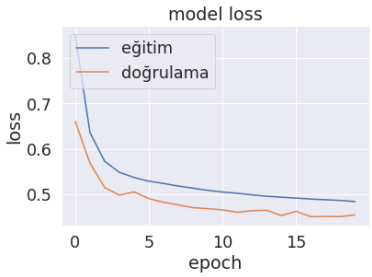
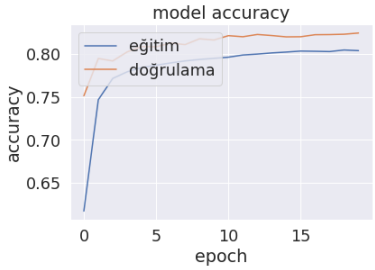
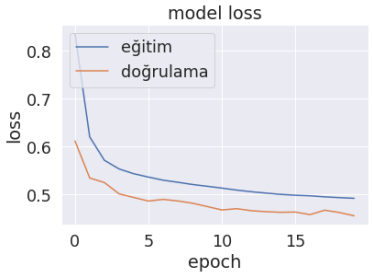
Tablo 5.2. (Devamı)

	Başarı Oranı	Kayıp Değeri
3		
4		
5		
6		
7		

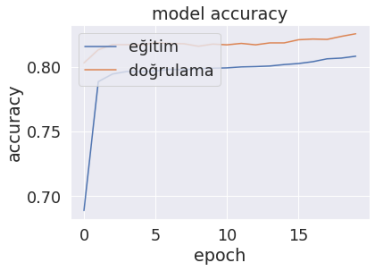
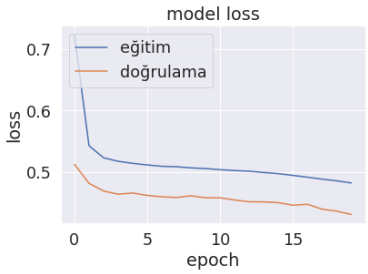
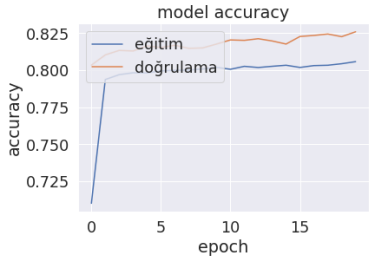
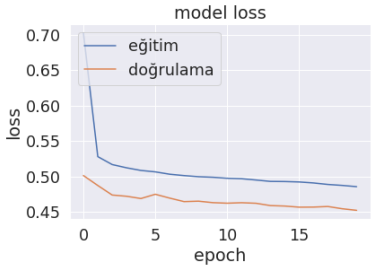
Tablo 5.3. Her bir doğrulama seti için LSTM modelinin eğitimi boyunca değişen başarı oranı ve kayıp fonksiyonu değeri grafikleri

	Başarı Oranı	Kayıp Değeri
1		
2		
3		
4		
5		

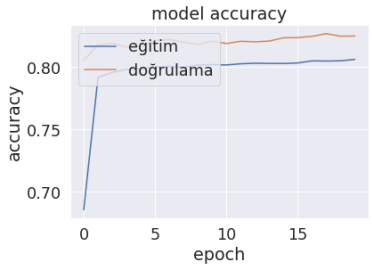
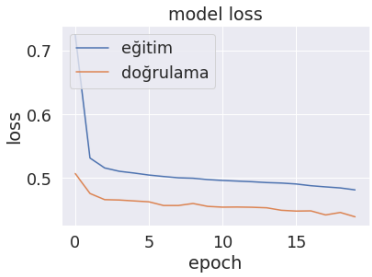
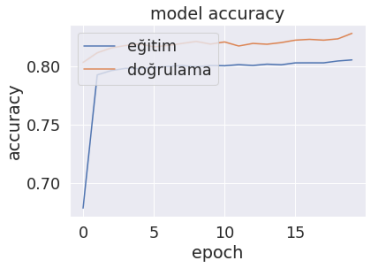
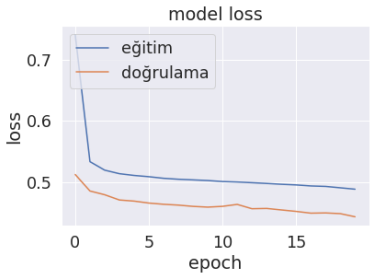
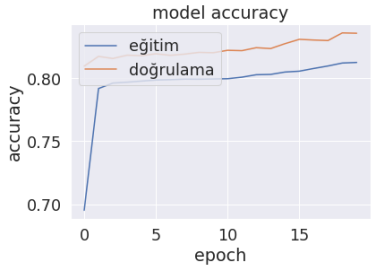
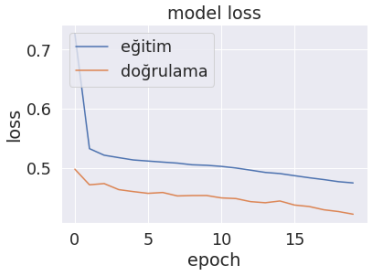
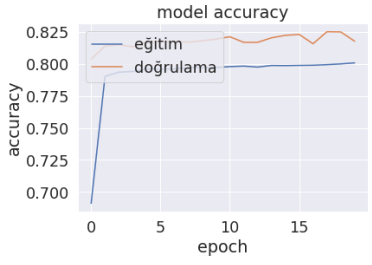
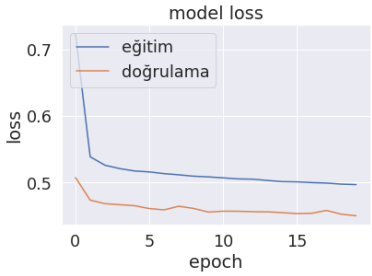
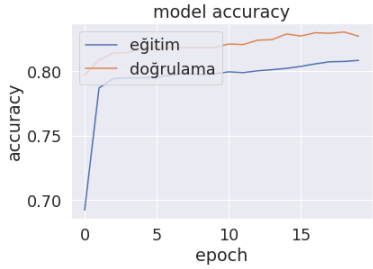
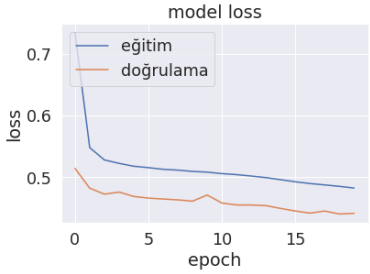
Tablo 5.3. (Devamı)

	Başarı Oranı	Kayıp Değeri
6		
7		

Tablo 5.4. Her bir doğrulama seti için GRU modelinin eğitimi boyunca değişen başarı oranı ve kayıp fonksiyonu değeri grafikleri

	Başarı Oranı	Kayıp Değeri
1		
2		

Tablo 5.4. (Devamı)

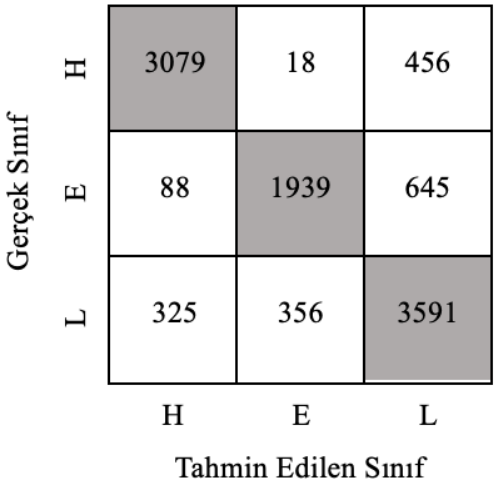
	Başarı Oranı	Kayıp Değeri
3		
4		
5		
6		
7		

Tablo 5.5.'te, CNN modelinin test sonuçları yer almaktadır. H, E ve L sınıfları için ortalama F skoru değerleri sırasıyla 0,86, 0,79 ve 0,81 olarak hesaplanmıştır. Sonuçlar incelendiğinde, 'H' olarak temsil edilen helix sınıfının diğer sınıflara göre daha başarılı tahmin edildiği görülmektedir. Modelin ortalama başarı oranı, 0,8254 olarak hesaplanmıştır.

Tablo 5.5. CNN modeli değerlendirme sonuçları

	Kesinlik			Duyarlılık			F skor			Başarı
	'H'	'E'	'L'	'H'	'E'	'L'	'H'	'E'	'L'	Oranı
1	0,88	0,84	0,77	0,87	0,73	0,84	0,87	0,78	0,80	0,8201
2	0,84	0,86	0,78	0,86	0,73	0,84	0,85	0,79	0,81	0,8164
3	0,89	0,83	0,77	0,84	0,76	0,85	0,86	0,79	0,81	0,8244
4	0,87	0,81	0,80	0,86	0,77	0,83	0,86	0,79	0,82	0,8261
5	0,87	0,82	0,79	0,85	0,74	0,84	0,86	0,78	0,81	0,8206
6	0,85	0,77	0,82	0,87	0,80	0,79	0,86	0,78	0,81	0,8206
7	0,91	0,84	0,79	0,87	0,78	0,86	0,89	0,81	0,82	0,8467
Ortalama	0,87	0,82	0,79	0,86	0,76	0,84	0,86	0,79	0,81	0,8254

Şekil 5.1.'de, CNN modelinin birinci çapraz doğrulama ile seti eğitim ve test edilmesinin sonucunda hesaplanan karmaşıklık matrisi gösterilmektedir. Matris incelendiğinde, test setinde bulunan 3553 adet 'H' sınıfı verisinin 3079'inin, 2672 adet 'E' sınıfı verisinin 1939'unun ve 4272 adet 'L' sınıfı verisinin 3591'inin doğru tahmin edildiği görülmektedir.



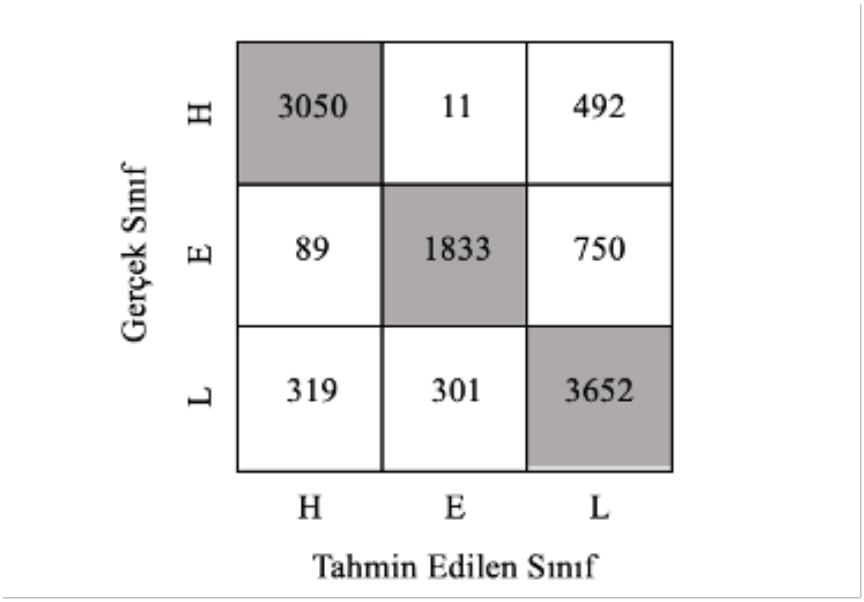
Şekil 5.1. CNN modeli birinci çapraz doğrulama seti karmaşıklık matrisi

Tablo 5.6.'da RNN modelinin her çapraz doğrulama kümesi için gerçekleştirilen test sonuçlarına göre hesaplanan değerlendirme metrikleri yer almaktadır. H, E ve L sınıfları için ortalama F skoru değerleri sırasıyla 0,86, 0,78 ve 0,81 olarak hesaplanmıştır. Tablo incelendiğinde, en iyi tahmin edilen sınıf 'H' ve modelin ortalama başarı oranı 0,8206 olduğu görülmektedir.

Tablo 5.6. RNN modeli değerlendirme sonuçları

	Kesinlik			Duyarlılık			F skor			Başarı
	'H'	'E'	'L'	'H'	'E'	'L'	'H'	'E'	'L'	Oranı
1	0,88	0,85	0,75	0,86	0,69	0,85	0,87	0,76	0,80	0,8131
2	0,84	0,81	0,80	0,86	0,79	0,80	0,85	0,80	0,80	0,8163
3	0,88	0,83	0,78	0,85	0,76	0,84	0,86	0,79	0,81	0,8227
4	0,89	0,84	0,77	0,83	0,72	0,87	0,86	0,77	0,82	0,8214
5	0,88	0,77	0,78	0,83	0,77	0,82	0,86	0,77	0,80	0,8129
6	0,88	0,78	0,80	0,85	0,76	0,82	0,86	0,77	0,81	0,8187
7	0,91	0,83	0,78	0,86	0,76	0,85	0,88	0,79	0,81	0,8365
Ortalama	0,88	0,82	0,78	0,85	0,75	0,84	0,86	0,78	0,81	0,8206

Şekil 5.2.'de, RNN modelinin birinci çapraz doğrulama seti ile eğitim ve test edilmesinin sonucunda hesaplanan karmaşıklık matrisi gösterilmektedir. Matris incelendiğinde, test setinde bulunan 3553 adet 'H' sınıfı verisinin 3050'sinin, 2672 adet 'E' sınıfı verisinin 1833'ünün ve 4272 adet 'L' sınıfı verisinin 3652'sinin doğru tahmin edildiği görülmektedir. Diğer çapraz doğrulama setleri ile yapılan testler sonucu hesaplanan karmaşıklık matrislerine Ekler bölümünde yer verilmiştir.



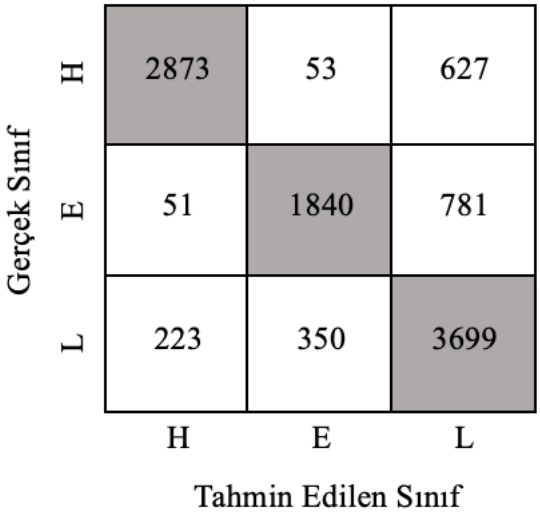
Şekil 5.2. RNN modeli birinci çapraz doğrulama seti karmaşıklık matrisi

LSTM modelinde ortalama başarı oranı 0,8110 olarak hesaplanırken geliştirilen modeller arasında en düşük performans gösteren model olmuştur. H, E ve L sınıfları için ortalama F skoru değerleri sırasıyla 0,86, 0,76 ve 0,80 olarak hesaplanmıştır. Diğer modellerde olduğu gibi, 'H' sınıfının tahmin oranı diğer sınıflara göre daha yüksek olduğu görülmektedir. (Tablo 5.7.)

Tablo 5.7. LSTM modeli değerlendirme sonuçları

	Kesinlik			Duyarlılık			F skor			Başarı
	‘H’	‘E’	‘L’	‘H’	‘E’	‘L’	‘H’	‘E’	‘L’	Oranı
1	0,91	0,82	0,72	0,81	0,69	0,87	0,86	0,75	0,79	0,8014
2	0,88	0,78	0,77	0,81	0,77	0,82	0,84	0,77	0,79	0,8022
3	0,89	0,81	0,75	0,82	0,73	0,84	0,85	0,77	0,80	0,8102
4	0,86	0,79	0,79	0,85	0,75	0,82	0,86	0,77	0,80	0,8130
5	0,86	0,76	0,79	0,84	0,77	0,80	0,85	0,76	0,79	0,8057
6	0,87	0,79	0,79	0,85	0,73	0,83	0,86	0,76	0,81	0,8150
7	0,90	0,84	0,76	0,86	0,71	0,85	0,88	0,77	0,80	0,8265
Ortalama	0,88	0,80	0,77	0,83	0,74	0,83	0,86	0,76	0,80	0,8110

Şekil 5.3.’te, LSTM modelinin üçüncü çapraz doğrulama seti ile eğitim ve test edilmesinin sonucunda hesaplanan karmaşıklık matrisi gösterilmektedir. Matris incelendiğinde, test setinde bulunan 3553 adet ‘H’ sınıfı verisinin 2873’ünün, 2672 adet ‘E’ sınıfı verisinin 1840’ının ve 4272 adet ‘L’ sınıfı verisinin 3699’unun doğru tahmin edildiği görülmektedir.



Şekil 5.3. LSTM modeli birinci çapraz doğrulama seti karmaşıklık matrisi

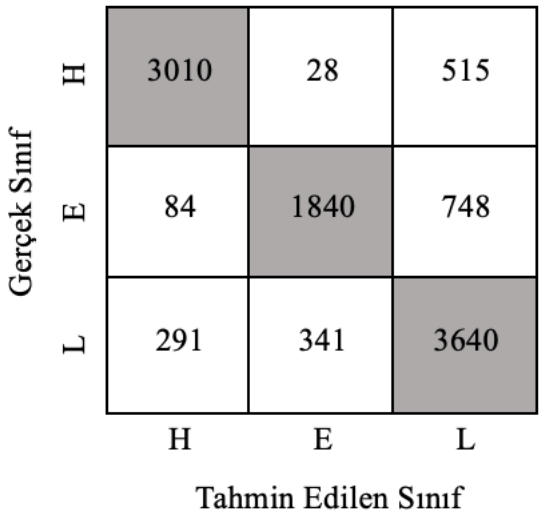
GRU modelinin ortalama başarı oranı, 0,8148 olarak hesaplanmıştır. H, E ve L sınıfları için ortalama kesinlik değerleri sırasıyla 0,86, 0,77 ve 0,80 olarak hesaplanmıştır. H

sınıfının başarılı tahmin edilme oranı diğer sınıflara göre daha yüksek olduğu görülmektedir. (Tablo 5.8.)

Tablo 5.8. GRU modeli değerlendirme sonuçları

	Kesinlik			Duyarlılık			F skor			Başarı
	'H'	'E'	'L'	'H'	'E'	'L'	'H'	'E'	'L'	Oranı
1	0,89	0,83	0,74	0,85	0,69	0,85	0,87	0,75	0,79	0,8088
2	0,87	0,80	0,77	0,83	0,74	0,83	0,85	0,77	0,80	0,8066
3	0,89	0,82	0,76	0,83	0,74	0,85	0,86	0,78	0,80	0,8167
4	0,87	0,80	0,78	0,84	0,74	0,83	0,86	0,77	0,80	0,8138
5	0,89	0,76	0,78	0,82	0,78	0,82	0,86	0,77	0,80	0,8125
6	0,87	0,77	0,79	0,85	0,73	0,82	0,86	0,75	0,80	0,8101
7	0,90	0,80	0,78	0,86	0,77	0,84	0,88	0,78	0,81	0,8327
Ortalama	0,88	0,80	0,77	0,84	0,74	0,83	0,86	0,77	0,80	0,8148

Şekil 5.4.'te, GRU modelinin birinci çapraz doğrulama seti ile eğitim ve test edilmesinin sonucunda hesaplanan karmaşıklık matrisi gösterilmektedir. Matris incelendiğinde, test setinde bulunan 3553 adet 'H' sınıfı verisinin 3010'unun, 2672 adet 'E' sınıfı verisinin 1840'ının ve 4272 adet 'L' sınıfı verisinin 3640'ının doğru tahmin edildiği görülmektedir.



Şekil 5.4. GRU modeli birinci çapraz doğrulama seti karmaşıklık matrisi

Tablo 5.9.'da modellerin toplam eğitim süreleri, ortalama F skoru ve ortalama başarı oranı ile başarı oranlarının standart sapma değerleri gösterilmektedir. Her çapraz doğrulama seti için eğitim süresi, CNN modelinde yaklaşık 1 dk. 57 sn., LSTM modelinde yaklaşık 2 dk. 48 sn., GRU modelinde yaklaşık 2 dk. 18 sn. sürmektedir. RNN modelinde eğitimin GPU ile gerçekleştirilmesi, CPU ile gerçekleştirilmesinden uzun sürdüğünden eğitim CPU kullanarak gerçekleştirilmiş ve bir çapraz doğrulama seti için yaklaşık 7 dk. 16 sn. sürmüştür.

Tablo 5.9. Modellerin performanslarının karşılaştırılması

	Eğitim Süresi	F skor	Başarı oranı	Standart Sapma
CNN	13 dk. 39 sn.	0,82	0,8254	0,0100
RNN	50 dk. 52 sn.	0,82	0,8206	0,0081
LSTM	19 dk. 36 sn.	0,81	0,8110	0,0087
GRU	16 dk. 18 sn.	0,81	0,8148	0,0087

BÖLÜM 6. SONUÇLAR VE ÖNERİLER

Bu tez çalışmasında, canlı organizmaların önemli bir parçası olan proteinlerin ikincil yapı tahmini için derin öğrenme yöntemleri kullanılmıştır. Protein yapısı, birincil yapıdan belirlenebilmesine rağmen yapının tamamını yalnızca birincil yapıdan tahmin etmek zordur. Bu sebeple, ikincil yapı tahmini proteinin üç boyutlu yapısını tahmin etmede önemli ve zor bir adımdır.

Bu çalışmada, proteinin birincil yapısından ikincil yapısının tahmini için derin öğrenme modellerinden CNN, RNN, LSTM ve GRU modelleri kullanılmıştır. Çalışma, Google Colaboratory ortamında belirtilen ağ modelleri ile CB513 veri seti kullanılarak gerçekleştirilmiştir. Test sonuçları başarı oranı üzerinden karşılaştırıldığında, CNN modeli %82,54 ile en başarılı derin öğrenme modeli olurken, %81,1 değeri ile LSTM en az başarılı derin öğrenme modeli olmuştur. F skorları üzerinden karşılaştırıldığında, CNN ve RNN modelleri %82 ile diğer iki modele göre %1 daha başarılı sonuç elde etmişlerdir.

GPU üzerinde gerçekleştirilen modeller arasında toplam eğitim süreleri incelendiğinde, 13 dk. 39 sn. ile CNN, en hızlı çalışan ağ olmuştur. LSTM ise 19 dk. 36 sn. ile en yavaş çalışan model olmuştur. RNN modeli ise yapılan test sonucunda, GPU üzerinde, CPU üzerinde çalıştığından daha yavaş çalıştığından CPU ile gerçekleştirilmiş ve toplam eğitim süresi 50 dk. 52 sn. de sonuç vermiştir.

Sonuç olarak, kullanılan modellerin başarı oranlarına bakıldığında birbirlerine yakın başarı oranları elde edilmiştir. Derin öğrenme yöntemleri ile protein ikincil yapı tahmini yapılırken bu çalışmada kullanılan dört modelin de kullanılabileceği görülmüştür. Bu alanda yapılan diğer çalışmalarda kullanılan geliştirme ortamına ve yöntemine bağlı olarak eğitim süreleri arasında büyük fark olduğu görülmüştür. Diğer

çalışmalarda eğitim günler sürerken yapılan çalışmada dakikalar bazında süre elde edilmiştir. Bu sebeple, hızlı sonuç almanın önemli olduğu protein ikincil yapı tahmini çalışmalarında, bu çalışmada gerçekleştirilen modellerin ve çalışma ortamının kullanılabileceği görülmektedir.

Derin öğrenme çalışmalarında, veri miktarının öğrenmeye etkisinin önemli olduğu bilinmektedir. Çalışmada önerilen modellerin başarısı veri miktarının arttırılması ile test edilebilir.

KAYNAKLAR

[1] Dill, K.A., MacCallum, J.L., The protein-folding problem, 50 years on. Science, 338(6110): 1042-1046, 2012.

[2] Bingöl, G., Proteinler, Ankara Üniversitesi Eczacılık Fakültesi Yayınları, 17-27, 1974.

[3] Liu H.L., Hsu, J.P., Recent developments in structural proteomics for protein structure determination. Proteomics, 5(8): 2056-2068, 2005.

[4] Branden, C., Tooze, J., Introduction to Protein Structure, 2. Baskı. Garland Science, New York, 3-12, 1999.

[5] Allison, L.A., From gene to protein. İçinde: Fundamental Molecular Biology. 1. Baskı, Blackwell Publishing, 79-107, 2007.

[6] http://80.251.40.59/veterinary.ankara.edu.tr/fidanci/Ders_Notlari/Ders_Notlari/Aminoasitler.html., Erişim Tarihi: 20.07.2021.

[7] https://eng.thesaurus.rusnano.com/wiki/article561., Erişim Tarihi: 20.07.2021

[8] Kabsch, W., Sander, C., Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22(12): 2577–2637, 1983.

[9] Pauling, L., Corey, R.B., Branson, H.R., The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. U. S. A., 37(4): 205-211, 1951.

[10] Pauling, L., Corey, R.B., The structure of synthetic polypeptides. Proc. Natl. Acad. Sci. U. S. A., 37(5): 241-250, 1951.

[11] Pauling, L., Corey, R.B., Atomic coordinates and structure factors for two helical configurations of polypeptide chains. Proc. Natl. Acad. Sci. U. S. A., 37(5): 235-240, 1951.

[12] Pauling, L., Corey, R.B., Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. Proc. Natl. Acad. Sci., 37(11): 729-740, 1951.

[13] Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., Phillips, D.C., A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature., 181(4610): 662-666, 1958.

[14] Perutz, M.F., Rossmann, M.G., Cullis, A.F., Muirhead, H., Will, G., North, A.C.T., Structure of Haemoglobin: A three-dimensional fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. Nature., 185(4711): 416-422, 1960.

[15] Rost, B., Sander, C., Third generation prediction of secondary structures, İçinde: Webster, D. (Ed.), Protein Struct. Predict. Methods Protoc., Humana Press, Clifton, NJ, 143: 71-95, 2000.

[16] Chou, P.Y., Fasman, G.D., Prediction of Protein Conformation. Biochemistry, 13(2): 222-245, 1974.

[17] Chou, P.Y., Fasman, G.D., Secondary structural prediction of proteins from their amino acid sequence. Trends Biochem. Sci., 2(6): 128-131, 1977.

[18] Garnier, J., Osguthorpe, D.J., Robson, B., Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol., 120(1): 97-120, 1978.

[19] Yi, T.M., Lander, E.S., Protein secondary structure prediction using nearest-neighbor methods. J. Mol. Biol., 232(4): 1117-1129, 1993.

[20] Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Lautrup, B., Nørskov, L., Olsen, O.H., Petersen, S.B., Protein secondary structure and homology by neural networks The α-helices in rhodopsin. FEBS Lett., 241(1-2): 223-228, 1988.

[21] Gibrat, J.F., Garnier, J., Robson, B., Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. J. Mol. Biol., 198(3): 425-443, 1987.

[22] Ptitsyn, O.B., Finkelstein, A. V., Theory of protein secondary structure and algorithm of its prediction. Biopolymers., 22(1): 15-25, 1983.

[23] Kneller, D.G., Cohen, F.E., Langridge, R., Improvements in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol., 214(1): 171-182, 1990.

[24] Rost, B., Sander, C., Combining evolutionary information and neural networks to predict protein secondary structure. Proteins Struct. Funct. Bioinforma., 9(1): 55-72, 1994.

[25] Kakumani, R., Devabhaktuni, V., Ahmad, M.O., A two-stage neural network based technique for protein secondary structure prediction. 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 1355-1358, 2008.

[26] Asai, K., Hayamizu, S., Handa, K., Prediction of protein secondary structure by the hidden Markov model. Bioinformatics, 9(2): 141-146, 1993.

[27] Aydin, Z., Altunbasak, Y., Borodovsky, M., Protein secondary structure prediction for a single-sequence using hidden semi-Markov models. BMC Bioinformatics, 7: 178, 2006.

[28] Martin, J., Gibrat, J.-F., Rodolphe, F., Hidden Markov Model for protein secondary structure. International Symposium on Applied Stochastic Models and Data Analysis, 2005.

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

[30] Nguyen, M.N., Rajapakse, J.C., Multi-class support vector machines for protein secondary structure prediction. Genome Inform., 14: 218-227, 2003.

[31] Chen, C., Tian, Y., Zou, X., Cai, P., Mo, J., Prediction of protein secondary structure content using support vector machine. Talanta, 71(5): 2069-2073, 2007.

[32] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17): 3389–3402, 1997.

[33] Remmert, M., Biegert, A., Hauser, A., Söding, J., HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods. 9(2): 173–175, 2012.

[34] Zhang, X., Mesirov, J.P., Waltz, D.L., Hybrid system for protein secondary structure prediction. J. Mol. Biol., 225(4): 1049-1063, 1992.

[35] Armano, G., Mancosu, G., Milanesi, L., Orro, A., Saba, M., Vargiu, E., A hybrid genetic-neural system for predicting protein secondary structure. BMC Bioinformatics., 6(4): 1-7, 2005.

[36] Min, S., Lee, B., Yoon, S., Deep learning in bioinformatics. Brief. Bioinform.,18(5): 851-869, 2017.

[37] Mamoshina, P., Vieira, A., Putin, E., Zhavoronkov, A., Applications of Deep Learning in Biomedicine. Mol. Pharm., 13(5):1445-1454, 2016.

[38] Baldi, P., Brunak, S., Frasconi, P., Soda, G., Pollastri, G., Exploiting the past and the future in protein secondary structure prediction, Bioinformatics, 15(11): 937-946, 1999.

[39] Pollastri, G., McLysaght, A., Porter: A new, accurate server for protein secondary structure prediction. Bioinformatics, 21(8): 1719-1720, 2005.

[40] Mirabello, C., Pollastri, G., Porter, PaleAle 4.0: High-accuracy prediction of protein secondary structure and relative solvent accessibility. Bioinformatics, 29(16): 2056-2058, 2013.

[41] Heffernan, R., Yang, Y., Paliwal, K., Zhou, Y., 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. 33(18): 2842-2849, 2017.

[42] Wang, S., Peng, J., Ma, J., Xu, J., Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Sci. Rep., 6: 18962, 2016.

[43] Wang, Y., Mao, H., Yi, Z., Protein secondary structure prediction by using deep learning method. Knowledge-Based Syst., 118: 115–123, 2017.

[44] Wardah, W., Khan, M.G.M., Sharma, A., Rashid, M.A., Protein secondary structure prediction using neural networks and deep learning: A review. Comput. Biol. Chem., 81: 1-8, 2019.

[45] Li, Z., Yu, Y., Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. IJCAI Int. Jt. Conf. Artif. Intell., New York City, 160-176, 2016.

[46] Guo, Y., Li, W., Wang, B., Liu, H., Zhou, D., DeepACLSTM: Deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction. BMC Bioinformatics, 20(1): 341, 2019.

[47] Wang, J., Cheng, J., Zhao, Z., Lu, W., Protein Secondary Structure Prediction Using Ensemble of LSTM Neural Networks. 2019 2nd Int. Conf. Inf. Syst. Comput. Aided Educ. (ICISCAE), Dalian, China, 241-244, 2019.

[48] Samuel, A.L., Some studies in machine learning using the game of checkers. IBM J. Res. Dev., 3(3): 210–229, 1959.

[49] Patterson, J., Gibson, A., Deep learning: A practitioner's approach, "O'Reilly Media, Inc.,", 251, 2017.

[50] Yani, M., Irawan, S., S.T., M.T., Application of Transfer Learning Using Convolutional Neural Network Method for Early Detection of Terry's Nail. J. Phys. Conf. Ser. IOP Publishing, 1201(1): 012052, 2019.

[51] https://towardsdatascience.com/the-most-intuitive-and-easiest-guide-for-convolutional-neural-network-3607be47480., Erişim Tarihi: 20.07.2021.

[52] ttps://colah.github.io/posts/2015-08-Understanding-LSTMs/., Erişim Tarihi: 20.07.2021

[53] Hochreiter, S., Schmidhuber, J., Long short-term memory. Neural Comput., 9(8): 1735-1780, 1997.

[54] Chung, J., Gulcehre, C., Cho, K., Bengio, Y., Empirical evaluation of gated recurrent neural networks on sequence modeling. ArXiv Prepr. ArXiv1412.3555, 2014.

[55] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15(1): 1929–1958, 2014.

[56] Kingma, D.P., Ba, J., Adam: A method for stochastic optimization. ArXiv Prepr. ArXiv1412.6980, 2014.

[57] Cuff, J.A., Barton, G.J., Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins Struct. Funct. Bioinforma. 34(4): 508-519, 1999.

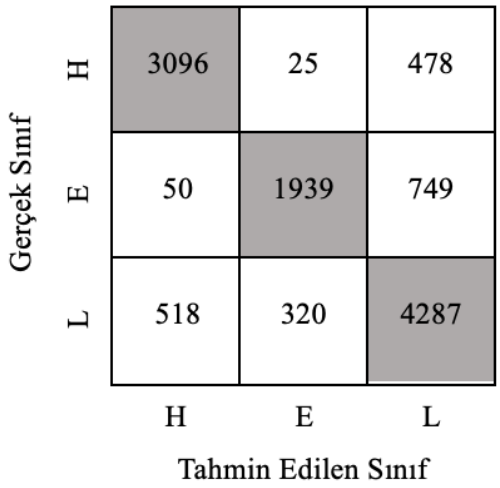
[58] Aydin, Z., Kaynar, O., Görmez, Y., Comparison of NR and UniClust databases for protein secondary structure prediction, IEEE Conferences 2018 26th Signal Processing and Communications Applications Conference (SIU), İzmir, 1-4, 2018.

[59] Aydin, Z., Protein İkincil Yapı Tahmini İçin Makine Öğrenmesi Yöntemlerinin Karşılaştırılması (Comparison of Machine Learning Classifiers for Protein Secondary Structure Prediction). IEEE Conferences 2018 26th Signal Processing and Communications Applications Conference (SIU), İzmir, 2018.

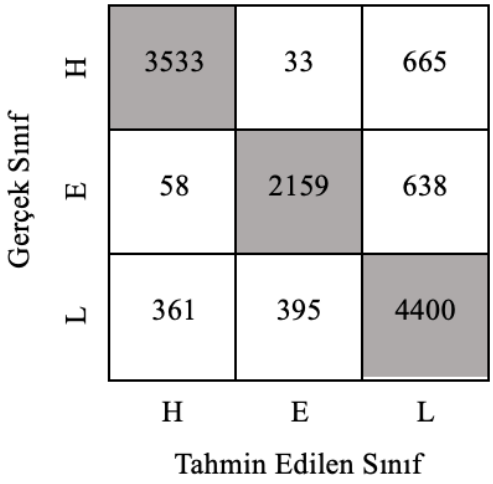
[60] https://research.google.com/colaboratory/intl/tr/faq.html., Erişim Tarihi: 20.07.2021.

EKLER

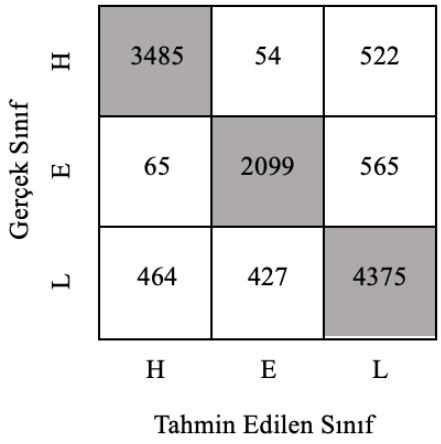
EK 1: CNN modeli ikinci çapraz doğrulama seti karmaşıklık matrisi



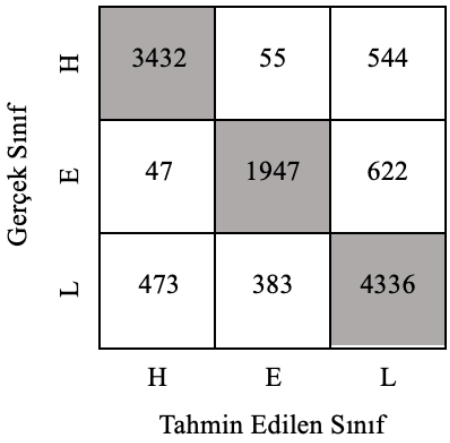
EK 2: CNN modeli üçüncü çapraz doğrulama seti karmaşıklık matrisi



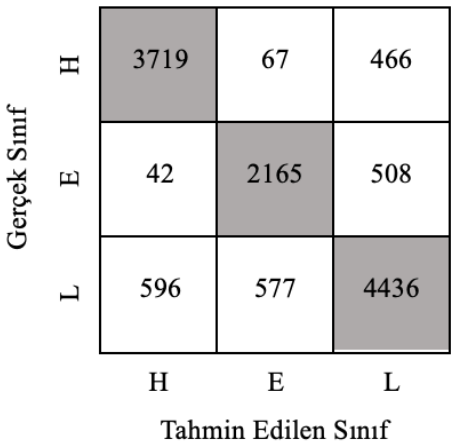
EK 3: CNN modeli dördüncü çapraz doğrulama seti karmaşıklık matrisi



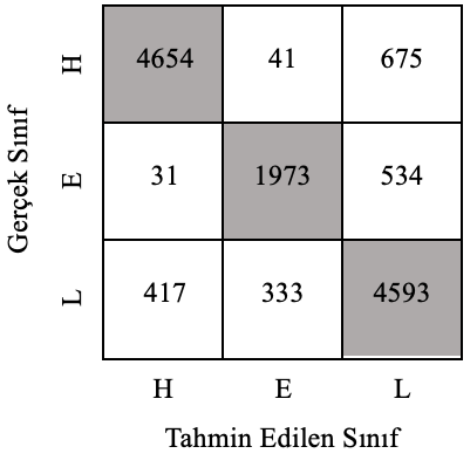
EK 4: CNN modeli beşinci çapraz doğrulama seti karmaşıklık matrisi



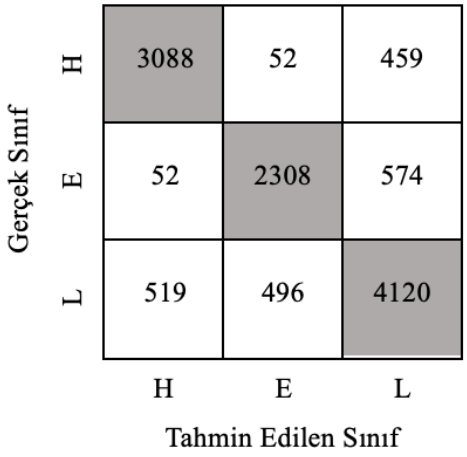
EK 5: CNN modeli altıncı çapraz doğrulama seti karmaşıklık matrisi



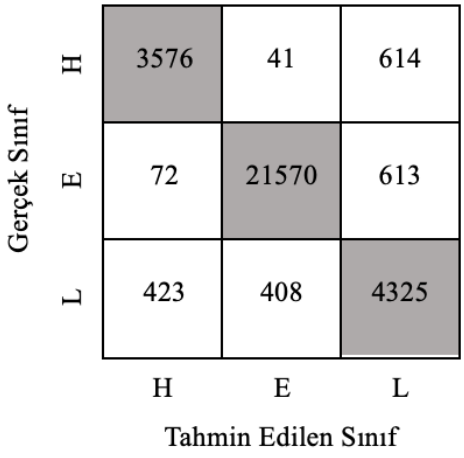
EK 6: CNN modeli yedinci çapraz doğrulama seti karmaşıklık matrisi



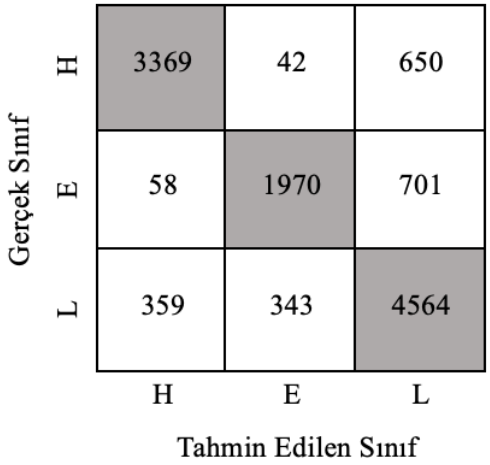
EK 7: RNN modeli ikinci çapraz doğrulama seti karmaşıklık matrisi



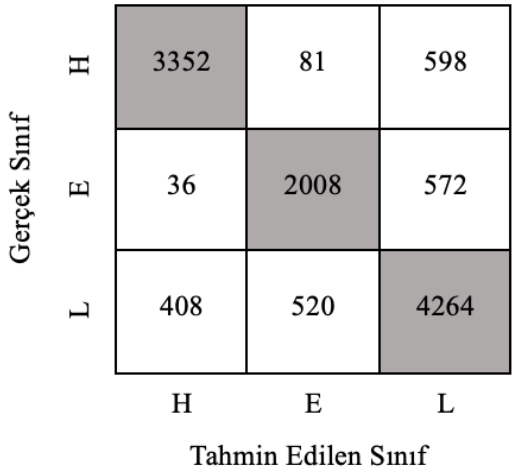
EK 8: RNN modeli üçüncü çapraz doğrulama seti karmaşıklık matrisi



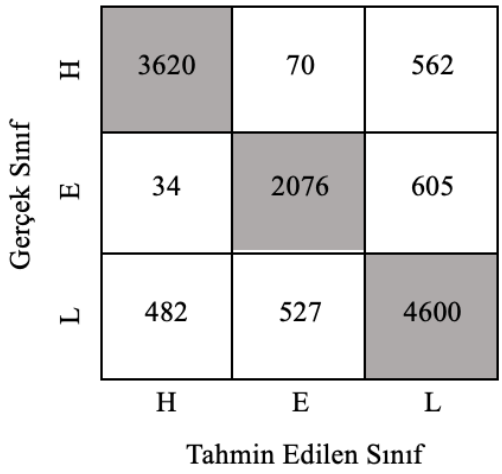
EK 9: RNN modeli dördüncü çapraz doğrulama seti karmaşıklık matrisi



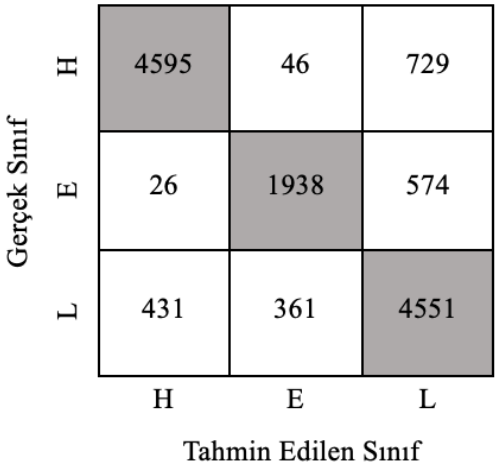
EK 10: RNN modeli beşinci çapraz doğrulama seti karmaşıklık matrisi



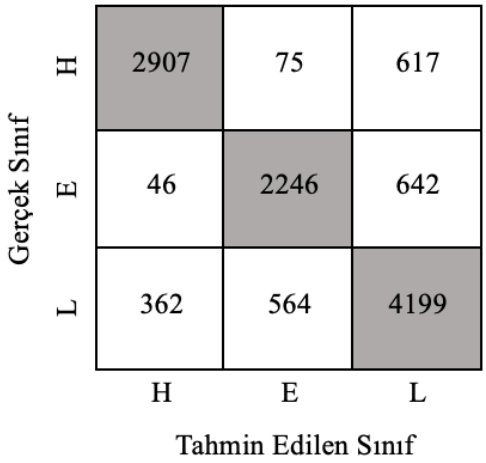
EK 11: RNN modeli altını çapraz doğrulama seti karmaşıklık matrisi



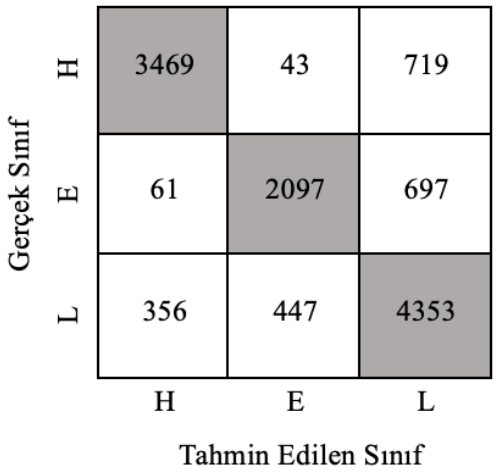
EK 12: RNN modeli yedinci çapraz doğrulama seti karmaşıklık matrisi



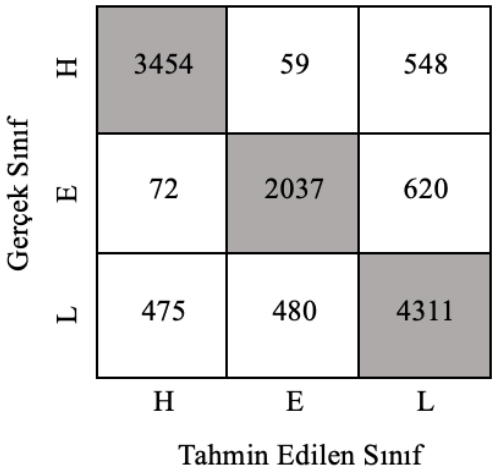
EK 13: LSTM modeli ikinci çapraz doğrulama seti karmaşıklık matrisi



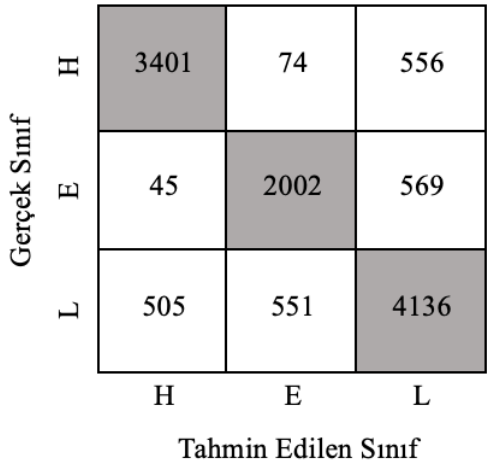
EK 14: LSTM modeli üçüncü çapraz doğrulama seti karmaşıklık matrisi



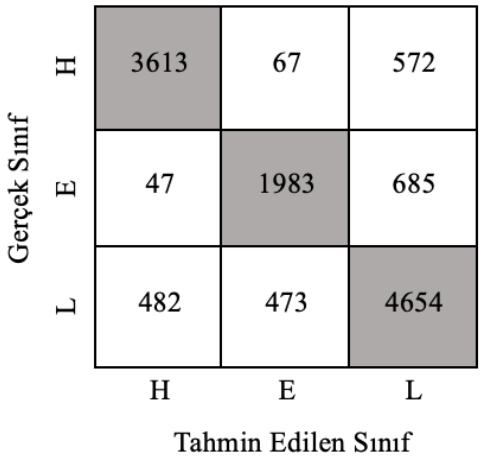
EK 15: LSTM modeli dördüncü çapraz doğrulama seti karmaşıklık matrisi



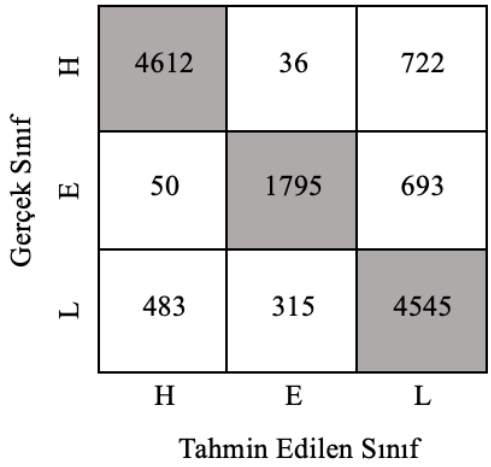
EK 16: LSTM modeli beşinci çapraz doğrulama seti karmaşıklık matrisi



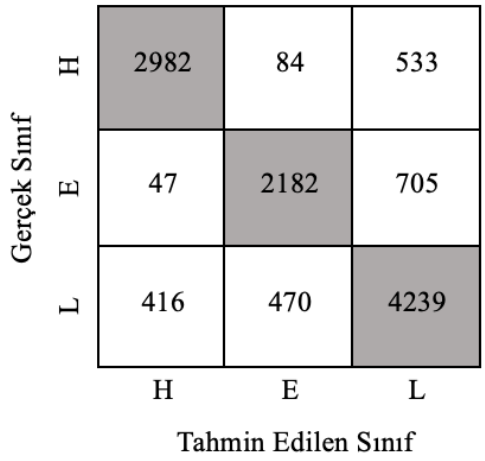
EK 17: LSTM modeli altıncı çapraz doğrulama seti karmaşıklık matrisi



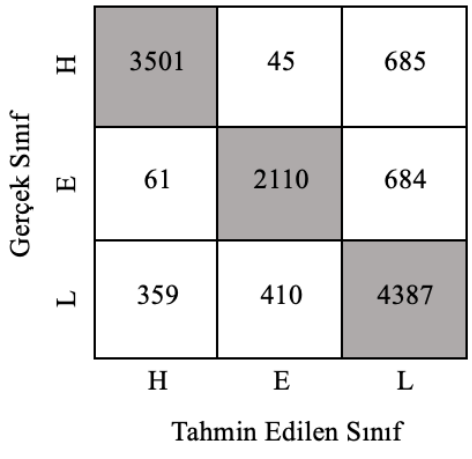
EK 18: LSTM modeli yedinci çapraz doğrulama seti karmaşıklık matrisi



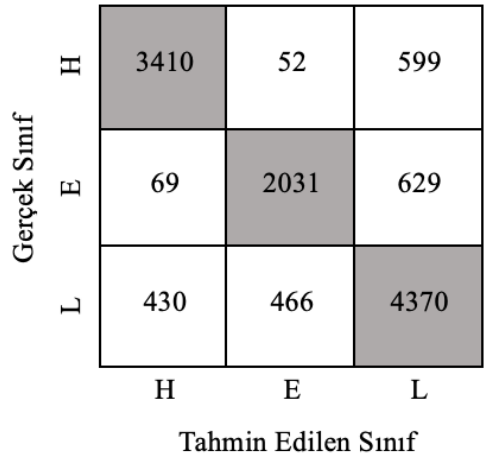
EK 19: GRU modeli ikinci çapraz doğrulama seti karmaşıklık matrisi



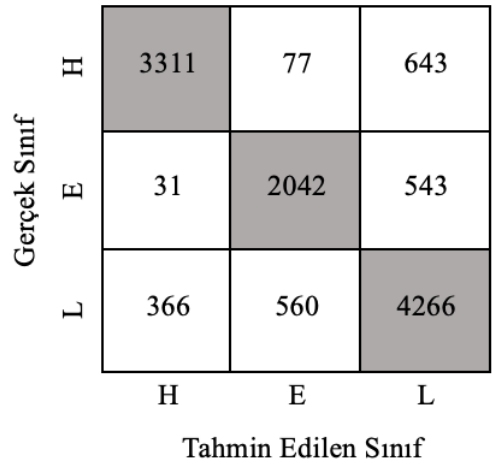
EK 20: GRU modeli üçüncü çapraz doğrulama seti karmaşıklık matrisi



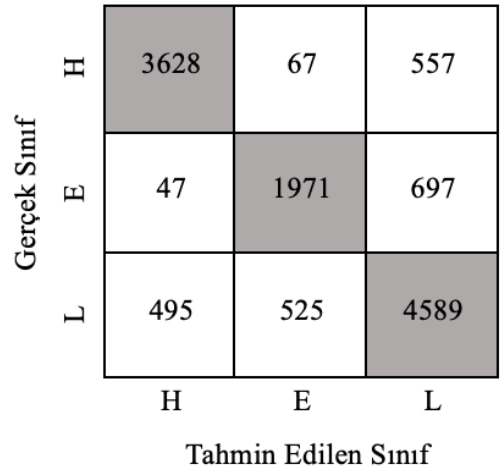
EK 21: GRU modeli dördüncü çapraz doğrulama seti karmaşıklık matrisi



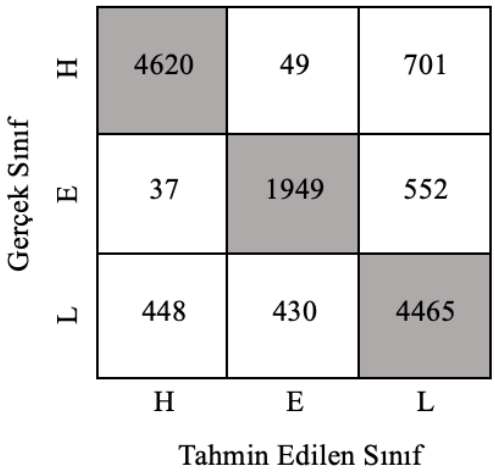
EK 22: GRU modeli beşinci çapraz doğrulama seti karmaşıklık matrisi



EK 23: GRU modeli altıncı çapraz doğrulama seti karmaşıklık matrisi



EK 24: GRU modeli yedinci çapraz doğrulama seti karmaşıklık matrisi



ÖZGEÇMİŞ

Adı Soyadı: Ezgi ÇAKMAK

ÖĞRENİM DURUMU

Derece	Eğitim Birimi	Mezuniyet Yılı
Yüksek Lisans	Sakarya Üniversitesi / Fen Bilimleri Enstitüsü / Bilişim Sistemleri Mühendisliği	Devam ediyor
Lisans	Sakarya Üniversitesi / Bilgisayar ve Bilişim Bilimleri Fakültesi / Bilişim Sistemleri Mühendisliği	2017
Lise	Nuri Nihat Aslanoba Anadolu Lisesi	2012

İŞ DENEYİMİ

Yıl	Yer	Görev
2021-Halen	Sakarya Üniversitesi	Kısmi Zamanlı Öğrenci

YABANCI DİL

İngilizce

ESERLER

1. CEDİMOĞLU İsmail Hakkı, SELVİ İhsan Hakan, ÇAKMAK Ezgi, Mühendislikte Yapay Zeka ve Uygulamaları 2: PANDAS ile Veri Analizi - Sakarya Üniversitesi - 2018