T.C.
KASTAMONU ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

BERT İLE İNSAN DNA SINIFLANDIRMA TAHMİNİ

ABDULQADER M.NOOR SAEED AL ALI KHAN

YÜKSEK LİSANS TEZİ

DOÇ. DR. EKMEL ÇETİN

HAZİRAN - 2026
KASTAMONU

TEZ ONAYI

ABDULQADER M.NOOR SAEED AL ALI KHAN tarafından hazırlanan "**BERT İLE İNSAN DNA SINIFLANDIRMA TAHMİNİ**" adlı tez çalışmasının savunma sınavı **03.06.2026** tarihinde yapılmış olup aşağıda verilen jüri tarafından oy birliği / oy çokluğu ile Kastamonu Üniversitesi Fen Bilimleri Enstitüsü **Bilgisayar Mühendisliği Ana Bilim Dalı Yüksek Lisans Tezi** olarak kabul edilmiştir.

Danışman	Doç. Dr. Ekmel ÇETİN Kastamonu Üniversitesi	.....
Jüri Üyesi	Dr. Öğr. Üyesi. Ali Burak ÖNCÜL Kastamonu Üniversitesi	.....
Jüri Üyesi	Dr. Öğr. Üyesi. Yusuf ALACA Hitit Üniversitesi	.....

Jüri üyeleri tarafından kabul edilmiş olan bu tez Kastamonu Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulunca onanmıştır.

Enstitü Müdürü Prof. Dr. Selahattin KAYMAKCI .....

TAAHHÜTNAME

Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırmalarının yapılması ve bulgularının analizlerinde bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu; ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını, bilimsel etiğe uygun olarak kaynak gösterildiğini bildirir ve taahhüt ederim.

ABDULQADER M.NOOR SAEED AL ALI KHAN

ÖZET

YÜKSEK LİSANS TEZİ

BERT İLE İNSAN DNA SIFINFLANDIRMA TAHMİNİ

ABDULQADER M.NOOR SAEED AL ALI KHAN

KASTAMONU ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI
DANIŞMAN:DOÇ. DR. EKMEL ÇETİN

Bu tez kapsamında, insan DNA dizilerinin sınıflandırılması amacıyla Transformer tabanlı BERT modeli kullanılmış ve söz konusu yaklaşımın etkinliği incelenmiştir. DNA dizileri, nükleotitlerin karmaşık dizilimleri, uzun menzilli bağımlılıklar ve biyolojik motiflerden oluşan bir yapıya sahiptir. Bu nedenle veri yapısı oldukça karmaşıktır. Geleneksel makine öğrenmesi yöntemleri, bu karmaşıklığı modellemede yeterli performansı sağlayamamaktadır. Bu doğrultuda, doğal dil işleme alanında başarılı sonuçlar veren BERT modeli, DNA dizilerini bir "biyolojik dil" olarak ele alacak şekilde uyarlanmıştır.

Çalışmada, 4380 DNA dizisinden oluşan ve yedi farklı sınıfa ayrılmış bir veri seti kullanılmıştır. DNA dizileri, 3-mer (üçlü nükleotit) temelli parçalara ayrılarak tokenizasyon işlemine tabi tutulmuştur. Tokenizasyon sürecinden geçirilen veriler, özel olarak geliştirilen bir PyTorch veri yükleyici aracılığıyla modele aktarılmıştır. Veri seti; eğitim için %80, doğrulama için %10 ve test için %10 olacak şekilde üç gruba ayrılmıştır. Eğitim sürecinde AdamW optimizasyon algoritması, çapraz entropi kaybı fonksiyonu, öğrenme oranı planlayıcısı ve gradyan kesme yöntemleri kullanılmıştır. Model performansı doğruluk, kesinlik, geri çağırma ve F1-skoru metrikleri ile değerlendirilmiştir. Elde edilen sonuçlar, BERT tabanlı modelin DNA dizilerindeki motifleri ve bağlamsal ilişkileri etkili biçimde öğrendiğini göstermektedir. Özellikle veri setinde daha az temsil edilen sınıflarda yanlış sınıflandırma oranlarının azaldığı gözlemlenmiştir. Ayrıca modelin dikkat mekanizmaları incelendiğinde, biyolojik açıdan anlamlı bölgelere odaklanabildiği belirlenmiştir.

Sonuç olarak, BERT tabanlı yaklaşımın DNA sınıflandırma probleminde geleneksel yöntemlerin ötesinde bir performans sunduğu görülmüştür. Elde edilen bulgular, genomik analizlerde Transformer tabanlı modellerin etkili biçimde kullanılabileceğini göstermekte ve biyoinformatik alanında gerçekleştirilecek yeni araştırmalar için önemli bir temel oluşturmaktadır.

ANAHTAR KELİMELER: DNA sınıflandırma, BERT, Transformer, derin öğrenme, biyoinformatik, k-mer tokenizasyon

Haziran 2026, 75 Sayfa

ABSTRACT

MSC THESIS

PREDICTING HUMAN DNA CLASSIFICATION USING BERT

ABDULQADER M.NOOR SAEED AL ALI KHAN

KASTAMONU UNIVERSITY INSTITUTE OF SCIENCE
DEPARTMENT OF COMPUTER ENGINEERING
SUPERVISOR:ASSOC. PROF. DR. EKMEL ÇETİN

This thesis investigates the effectiveness of a Transformer-based BERT model for the classification of human DNA sequences. DNA sequences consist of complex arrangements of nucleotides, long-range dependencies, and biologically meaningful motifs, making their structure highly intricate. Traditional machine learning approaches often fail to adequately capture this complexity. In this context, the BERT model—originally successful in natural language processing—has been adapted to treat DNA sequences as a form of "biological language."

The study utilizes a dataset comprising 4,380 DNA sequences categorized into seven distinct classes. The sequences were tokenized using a 3-mer (trinucleotide) segmentation approach. The tokenized data were then fed into the model באמצעות a custom-developed PyTorch data loader. The dataset was divided into three subsets: 80% for training, 10% for validation, and 10% for testing. During training, the AdamW optimization algorithm, cross-entropy loss function, learning rate scheduler, and gradient clipping techniques were employed. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics. The results demonstrate that the BERT-based model effectively learns motifs and contextual relationships within DNA sequences. Notably, a reduction in misclassification rates was observed, particularly for underrepresented classes. Furthermore, analysis of the model's attention mechanisms revealed that it can focus on biologically meaningful regions.

In conclusion, the BERT-based approach outperforms traditional methods in DNA classification tasks. The findings indicate that Transformer-based models can be effectively utilized in genomic analysis and provide a strong foundation for future research in the field of bioinformatics.

KEYWORDS: DNA classification, BERT, Transformer, deep learning, bioinformatics, k-mer tokenization

July 2026, 75 Page

TEŞEKKÜR

Tezimi hazırlarken her aşamada yanımda olan, bilgi ve deneyimiyle bana yol gösteren değerli danışmanım Doç. Dr. Ekmel ÇETİN'e gönülden teşekkür ediyorum. Onun bilimsel yaklaşımı, sabrı ve desteği, çalışmamın şekillenmesinde gerçekten çok etkili oldu. Eğitim hayatım boyunca bana inanan ve bugünlere gelmemde emeği büyük olan aileme, moral ve motivasyonlarıyla her zaman destek olan arkadaşlarıma da ayrıca teşekkür etmek istiyorum. Çalışmam sırasında doğrudan ya da dolaylı şekilde katkı sunan tüm akademisyenlere, araştırmacılara ve kaynaklarını paylaşarak bilime destek olan herkese minnettarım.

ABDULQADER M.NOOR SAEED AL ALI KHAN

Kastamonu, 2026

İÇİNDEKİLER

Sayfa

TEZ ONAYI ..... **ii**
TAAHHÜTNAME ..... **iii**
ÖZET ..... **iv**
ABSTRACT ..... **v**
TEŞEKKÜR ..... **vi**
İÇİNDEKİLER ..... **vii**
ŞEKİLLER DİZİNİ ..... **ix**
TABLOLAR DİZİNİ ..... **x**
SİMGELER VE KISALTMALAR DİZİNİ ..... **xi**
1. GİRİŞ **1**
1.2. Literatürdeki Boşluklar ve Bu Tezin Katkıları 10
1.2.1. DNA'ya Özel Tokenizasyon Eksikliği ve Geliştirilmesi 11
1.2.2. Tam Genom Ölçeğinde Analiz ve Uzun Menzilli Etkileşimlerin Modellenmesi Sınırlılığı 11
1.2.3. Yorumlanabilirlik Sorunu ve Biyolojik İçgörülerin Elde Edilmesi 12
1.3. Sınırlamalar 12
2. LİTERATÜR ARAŞTIRMASI **14**
2.1. DNA (Deoksiribonükleik Asit) 14
2.2. DNA Sınıflandırmasında Geleneksel Yöntemler ve Sınırlamaları 17
2.2.1. k-mer Tabanlı Yaklaşımlar 18
2.2.2. Saklı Markov Modelleri (HMM) 18
2.2.3. Geleneksel Yöntemlerin Başlıca Sınırlamaları 19
2.3. Derin Öğrenme Tabanlı DNA Sınıflandırma Yöntemleri 21
2.3.1. Evrişimli Sinir Ağları (CNN) 21
2.3.2. Tekrarlayan Sinir Ağları (RNN/LSTM) 22
2.3.3. Hibrit Modeller (CNN-BiLSTM, CNN-Attention) 22
2.3.4. Derin Öğrenmenin Sınırlamaları 23
2.4. Dönüştürücü (Transformer) Tabanlı Modeller ve BERT'in Uygulanması 24
3. MATERYALLER VE YÖNTEMLER **29**
3.1. Veri Seti 29
3.1.1. Veri Kaynağı 29
3.1.2. Veri Türü 29
3.1.3. Sınıf Sayısı 30
3.1.4. Veri Seti Bölme 30
3.1.5. Sınıf Etiketlerinin Biyolojik Fonksiyonları 31
3.2. Ön İşleme ve Tokenizasyon 31
3.2.1. K-mer Tokenizasyonu ve Sözlük Oluşturma 32
3.2.2. Özel Veri Seti Sınıfı (DNADataset) 32
3.2.3. Collate Fonksiyonu (collate_fn) 33
3.3. Model Mimarisi 34
3.3.1. BERT'in Dikkat Mekanizması 34
3.3.2. Katman Normalizasyonu 35
3.3.3. BertClassifier Sınıfı 35
3.3.4. BERT Konfigürasyonu 35

3.3.5. Bert For Sequence Classification 37
3.3.6. Model Özeti 37
3.4. Eğitim Prosedürü 37
3.4.1. Optimizasyon 38
3.4.2. Kayıp Fonksiyonu 38
3.4.3. Öğrenme Oranı Planlayıcısı (Scheduler) 39
3.4.4. Gradyan Kesme (Gradient Clipping) 39
3.4.5. Eğitim Döngüsü 39
3.4.6. Model Kaydetme 40
3.5. Değerlendirme Metrikleri ve Görselleştirme 40
3.5.1. Metrikler 40
3.6. Test Seti Üzerinde Nihai Değerlendirme 42
3.7. Tek Bir Dizi İçin Tahmin 42
3.8. Yazılım ve Donanım Ortamı 43
4. BULGULAR 45
4.1. Eğitim Süreci Analizi 45
4.1.1. Modelin Eğitim ve Doğrulama Kaybı Grafiği 45
4.1.2. Modelin Eğitim ve Doğrulama Doğruluk Grafiği 46
4.1.3. Modelin Eğitim ve Doğrulama Hassasiyet ve Geri Çağırma Grafiği 48
4.1.4. Modelin Doğrulama F1-Skor Grafiği 48
4.2. Test Seti Üzerinde Nihai Değerlendirme 49
4.2.1. Sınıflandırma Raporu 49
4.2.2. Karışıklık Matrisi 50
4.3. Tek Bir Dizi İçin Tahmin Örneği 51
4.4. Literatürle Karşılaştırma 51
5. SONUÇ VE TARTIŞMA 54
5.1. Gelecek Çalışmalar İçin Öneriler 56
KAYNAKLAR 57
ÖZGEÇMİŞ 64

ŞEKİLLER DİZİNİ

Sayfa

Şekil 1.1 DNA çift sarmal yapısı. 2
Şekil 1.2 DNA Yapısı 2
Şekil 1.3 BERT mimarisinin basitleştirilmiş görünümü 6
Şekil 1.4 BERT tabanlı DNA sınıflandırma süreci 10
Şekil 2.1 DNA'nın Çift Sarmal Yapısı ve Baz Bileşenleri. 15
Şekil 2.2 gkm-SVM vs. k-mer-SVM performans grafiği. 20
Şekil 2.3 k-mer histogram + CNN sınıflandırma akış diyagramı 20
Şekil 2.4 PC-mer encodlama ve sınıflandırma süreci 21
Şekil 2.5 Genel CNN-LSTM mimari şeması 24
Şekil 2.6 DNA-BERT tabanlı dönüştürücü derin öğrenme modelinin eğitim süreci 27
Şekil 2.7 Transformer kodlayıcı katmanı 27
Şekil 3.1 human.txt veri setinden örnek DNA görünümü 31
Şekil 3.2 Tokenizasyon ve ön işleme örneği 34
Şekil 4.1 Modelin eğitim ve doğrulama kaybı grafiği 45
Şekil 4.2 Modelin eğitim ve doğrulama doğruluk grafiği 46
Şekil 4.3 Modelin eğitim ve doğrulama hassasiyet ve geri çağırma grafiği 48
Şekil 4.4 Modelin Doğrulama F1-Skor gerafiği 48
Şekil 4.5 Test Seti Üzerindeki Karışıklık Matrisi 50
Şekil 4.6 Tek bir dizi için tahmin örneği 51

TABLOLAR DİZİNİ

Sayfa

Tablo 4.1 Test Seti Üzerindeki Sınıflandırma Raporu 49
Tablo 4.2 Test Seti Performans Karşılaştırması 51
Tablo 4.3 Aynı Veri Seti Üzerinde CNN-LSTM ve BERT Sınıflandırma Karşılaştırması 51

SİMGELER VE KISALTMALAR DİZİNİ

Simgeler

A : Adenin
T : Timin
C : Sitozin
G : Guanin
N : Belirsiz/ bilinmeyen nükleotit
k-mer : k uzunluğundaki alt dizi (ör. 3-mer)
H : Entropi değeri
L : Dizi uzunluğu
α : Öğrenme oranı katsayısı
θ : Model parametreleri

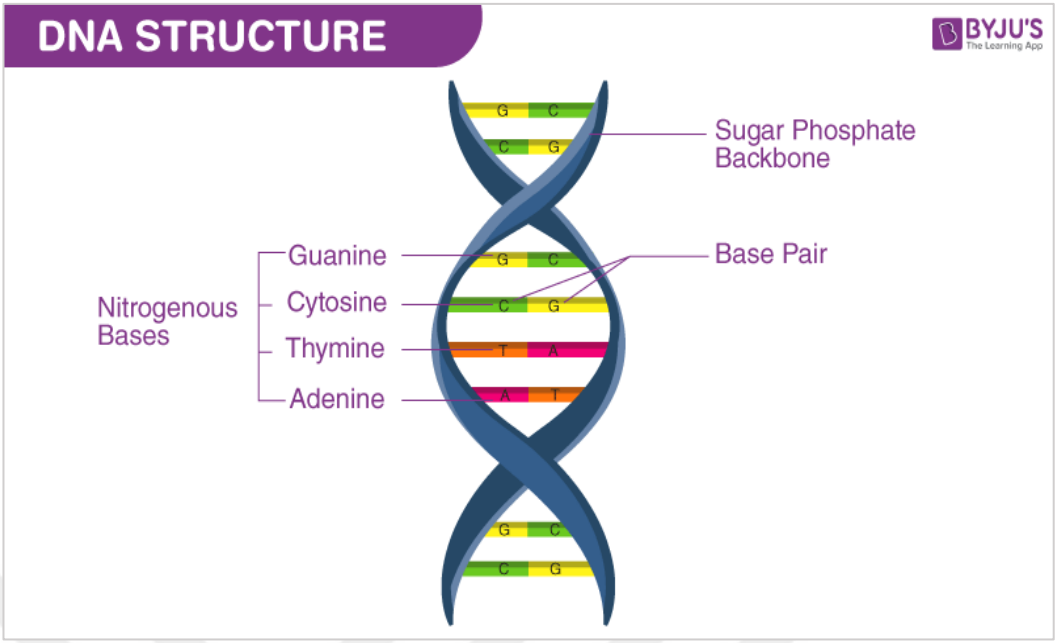
Kısaltmalar

DNA : Deoksiribonükleik Asit
RNA : Ribonükleik Asit
BERT : Bidirectional Encoder Representations from Transformers
NLP : Natural Language Processing (Doğal Dil İşleme)
CNN : Convolutional Neural Network (Evrişimli Sinir Ağı)
RNN : Recurrent Neural Network
LSTM : Long Short-Term Memory
BiLSTM : Bidirectional Long Short-Term Memory
HMM : Hidden Markov Model
GPU : Grafik İşlem Birimi
TPU : Tensor Processing Unit
SVM : Support Vector Machine
AUC : Area Under the Curve
F1 : F1-Skoru
TF : Transkripsiyon faktörü
ML : Makine Öğrenimi
DL : Derin Öğrenme

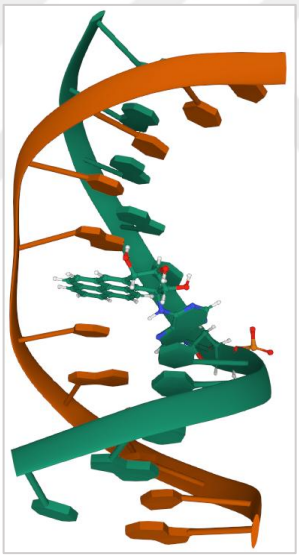
1. GİRİŞ

Biyoinformatik, farklı disiplinlerin kesişiminde yer alan bir araştırma alanıdır. Biyolojik verilerin toplanması, saklanması, analiz edilmesi ve yorumlanması süreçlerinde bilgisayar teknolojilerinden yararlanılmaktadır (Lesk, 2013; Mount, 2004). Özellikle genetik diziler gibi yüksek karmaşıklığa sahip verilerin incelenmesinde biyoinformatik merkezi bir rol üstlenmektedir. Deoksiribonükleik asit (DNA), canlı organizmaların ve bazı virüslerin genetik bilgisini taşıyan, yaşamın temelini oluşturan kritik bir moleküldür. Genetik bilginin aktarımı, protein sentezi ve evrimsel değişimlerin ortaya çıkması süreçlerinde temel bir işlev görmektedir. Bu nedenle DNA, biyolojik sistemlerin merkezinde yer almaktadır (Yang vd., 2024).

DNA molekülü, adenin (A), timin (T), guanin (G) ve sitozin (C) olmak üzere dört farklı nükleotitten oluşmaktadır. Şeker-fosfat omurgasına bağlı olan bu nükleotitler çift sarmal yapı şeklinde düzenlenmektedir. Adenin–timin ve guanin–sitozin baz eşleşmeleri, çift sarmal yapının stabilitesini sağlayarak genetik bilginin doğru biçimde kopyalanmasına ve aktarılmasına olanak tanımaktadır (Cabral-de-Mello vd., 2024).



Şekil 1.1 DNA çift sarmal yapısı (Byjus.com, n.d.).



Şekil 1.2 DNA Yapısı (Öncül, 2022)

Şekil 1.1 ve Şekil 1.2 görsellerde, DNA'nın çift sarmal yapısını, nükleotidleri (A, T, G, C), şeker-fosfat omurgasını ve baz eşleşmelerini (A-T, G-C) açıkça göstermelidir.

DNA'nın temel işlevleri arasında, hücrelerin faaliyetlerini ve organizmanın gelişimini yönlendiren genetik bilgiyi taşıması yer almaktadır. Bu genetik bilgi, transkripsiyon süreci aracılığıyla RNA moleküllerine aktarılmakta ve translasyon sürecinde protein sentezinde kullanılmaktadır. Söz konusu mekanizmalar, gen ekspresyonunun temelini oluşturmaktadır. DNA'da meydana gelen mutasyonlar, organizmaların genetik

çeşitliliğini artırmakta; bu durum evrimsel süreçlerde önemli bir rol oynayarak canlıların çevresel koşullara uyum sağlamasına katkıda bulunmaktadır.

DNA'nın keşfedilmesi ve yapısının aydınlatılmasının ardından genetik mühendisliği alanı gelişmiş, CRISPR-Cas9 gibi genom düzenleme teknolojileri ortaya çıkmıştır. Bu gelişmeler modern biyoteknolojinin ilerlemesine önemli katkılar sağlamıştır (Smith ve Brown, 2019). Sonuç olarak, söz konusu teknolojik ilerlemeler tıp, tarım ve çevre bilimleri başta olmak üzere birçok alanda devrim niteliğinde gelişmelerin yaşanmasına olanak tanımıştır (Chen ve Zhang, 2020; Yamamoto vd., 2024).

Günümüzde DNA üzerine yürütülen araştırmalar yalnızca temel bilim çalışmalarıyla sınırlı kalmamakta, aynı zamanda uygulamalı alanlarda da önemli bir yer tutmaktadır. Genomik, metagenomik ve filogenetik gibi araştırma alanları giderek daha fazla önem kazanmıştır. Genom kavramı, bir organizmaya ait tüm genetik bilginin bütününü ifade etmektedir. Metagenomik, çevresel örneklerden elde edilen DNA verileri aracılığıyla mikrobiyal toplulukların incelenmesini sağlamaktadır. Filogenetik ise türler arasındaki evrimsel ilişkilerin ortaya konulmasına katkı sunmaktadır. Özellikle son yıllarda DNA'nın yapısı ve işleyişine ilişkin bilgi birikiminin artmasıyla birlikte, bu alanlarda önemli bilimsel ilerlemeler kaydedilmiştir (Minamoto, 2022).

DNA, hücresel süreçlerde hayati bir rol üstlenmekle birlikte işlevlerini tek başına gerçekleştirmemektedir. Bu süreçlerde, binlerce DNA bağlayıcı protein (DNA-binding proteins, DBP) görev almaktadır. Söz konusu proteinler olmadan DNA replikasyonu ve rekombinasyonu gibi temel biyolojik işlevlerin gerçekleşmesi mümkün değildir. Bununla birlikte, DBP'lerin tanımlanması ve sınıflandırılması oldukça zorlu bir süreçtir (Qu vd., 2019).

Son yıllarda araştırmacılar, DNA bağlayıcı proteinlerin sınıflandırılmasında geleneksel makine öğrenmesi (ML) yöntemlerinden yararlanmaktadır. Makine öğrenmesi algoritmalarının, özellikle DBP tanımlama görevlerinde başarılı sonuçlar sunduğu gösterilmiştir (Liu vd., 2015). Diğer yöntemlerle karşılaştırıldığında, geleneksel ML yaklaşımlarının daha düşük maliyetli, daha hızlı ve büyük veri setlerini yönetme açısından daha avantajlı olduğu görülmektedir (Arıcan, 2023). Dizi tabanlı

tahmin yöntemleri ise proteinlerin sekans bilgisi üzerinden çalışabilmeleri ve ek yapısal bilgi gerektirmemeleri nedeniyle DBP tanımlama çalışmalarında önemli bir avantaj sağlamaktadır.

Makine öğrenmesine dayalı DBP tanımlama yöntemlerinde, özellik temsil yöntemleri ile sınıflandırma algoritmaları temel bileşenleri oluşturmaktadır. Bir proteinin sayısal özellikler aracılığıyla temsil edilmesi, biyoinformatik analizlerin etkin biçimde gerçekleştirilmesine olanak tanımaktadır. ML tabanlı özellik temsil yaklaşımları genel olarak iki ana grupta ele alınmaktadır: yapı tabanlı tahmin yöntemleri ve dizi tabanlı tahmin yöntemleri.

Son yıllarda derin öğrenme (Deep Learning, DL) modelleri, özellik çıkarımı alanında önemli ilerlemeler sağlamıştır (Jurtz vd., 2017). DNA dizilerinin sınıflandırılmasında Evrişimsel Sinir Ağları (Convolutional Neural Networks, CNN), CNN-LSTM ve CNN-Bi-LSTM gibi modeller yaygın biçimde kullanılmaktadır (Rizzo vd., 2016). Bu derin öğrenme mimarileri, ham verilerden anlamlı özellikleri otomatik olarak öğrenebilmekte ve elde edilen temsilleri sınıflandırma süreçlerinde etkin şekilde kullanabilmektedir. Özellikle büyük ve karmaşık veri setlerinde, manuel özellik çıkarımına ihtiyaç duymadan öğrenme gerçekleştirebilmeleri nedeniyle önemli avantajlar sunmaktadır. Bu durum, DNA ve protein sınıflandırma problemlerinde derin öğrenme yöntemlerini güçlü bir alternatif hâline getirmektedir.

Protein sınıflandırmasına yönelik olarak dizi verilerine dayalı çeşitli derin öğrenme modelleri geliştirilmiştir. Örneğin DeepDBP modeli, DNA bağlayıcı proteinlerin belirlenmesi amacıyla 12 katmanlı bir derin sinir ağı mimarisi kullanmaktadır (Shadab vd., 2020). iDNAProt-ES modeli ise amino asit dizilerinin evrimsel özelliklerini dikkate alarak destek vektör makineleri aracılığıyla sınıflandırma gerçekleştirmektedir (Hu vd., 2019). Derin öğrenme yaklaşımları yalnızca protein ailelerinin hiyerarşik sınıflandırılmasında değil, farklı biyoinformatik uygulamalarda da kullanılmaktadır. DeepFam modeli, protein dizilerini aile, alt aile ve daha alt seviyelerde sınıflandırabilme yeteneğine sahiptir (Sandaruwan ve Wannige, 2021). Bunun yanı sıra, derin öğrenme yöntemleri DNA transkripsiyon faktörlerinin sınıflandırılmasında da uygulanmaktadır (Öncül, 2022). DeepCAC modeli, evrişimsel katmanları çoklu

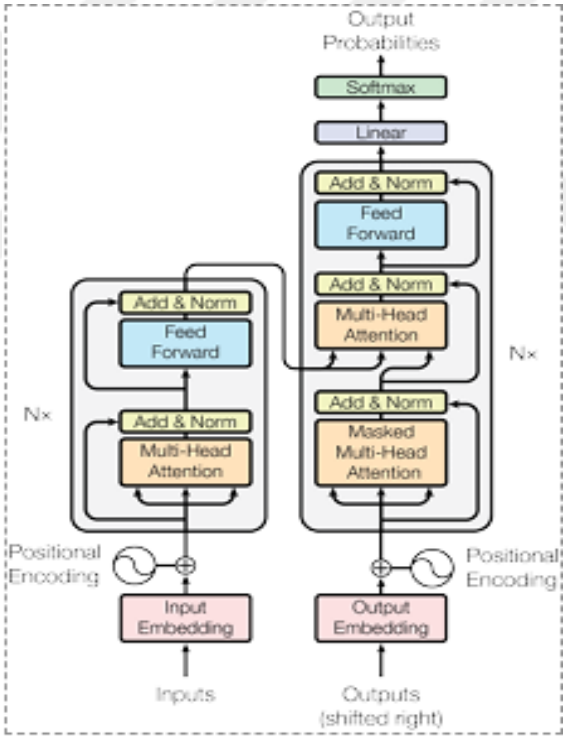
başlı öz-dikkat mekanizmasıyla birleştirerek DNA transkripsiyon faktörlerinin sınıflandırılmasını gerçekleştirmektedir. Ayrıca, zayıf denetimli öğrenme yaklaşımı ile çoklu örnek öğrenmenin birleştiği hibrit derin sinir ağı modelleri kullanılarak in-vivo protein-DNA bağlanma süreçlerinin modellenmesine yönelik çalışmalar da bulunmaktadır (Zhang vd., 2019).

Genel olarak değerlendirildiğinde, derin öğrenme modelleri DNA ve protein sınıflandırma problemlerinde önemli bir potansiyel ortaya koymaktadır. Bu modeller, protein fonksiyonlarının tahmin edilmesi ve dizi temelli analizler aracılığıyla proteinlerin farklı aile ve alt ailelere ayrılması süreçlerinde yüksek doğruluk ve etkinlik sağlamaktadır. Elde edilen bulgular, derin öğrenme yaklaşımlarının biyoinformatik ve hesaplamalı biyoloji alanlarında önemli bir etki oluşturabileceğini göstermektedir.

Bu tez kapsamında, biyoinformatik alanında DNA dizilerinin sınıflandırılması amacıyla derin öğrenme tabanlı modellerin, özellikle Transformer mimarisine dayanan BERT modelinin etkinliği incelenmektedir. BERT modeli, Google tarafından geliştirilmiş olup doğal dil işleme alanında önemli bir dönüşüm yaratmıştır (Devlin vd., 2018). BERT, "Bidirectional Encoder Representations from Transformers" ifadesinin kısaltmasıdır ve modelin temel özelliği bağlamı iki yönlü olarak değerlendirebilmesidir. Geleneksel dil modellerinin aksine yalnızca kelimenin öncesine ya da sonrasına odaklanmak yerine, metnin tamamını eş zamanlı olarak analiz etmektedir. Bu sayede model, bağlama bağlı anlam farklılıklarını ayırt edebilmekte ve aynı kelimenin farklı kullanım durumlarını doğru biçimde yorumlayabilmektedir.

BERT modeli, arama motorları, makine çevirisi sistemleri, soru-cevap uygulamaları ve konuşma tabanlı yapay zekâ sistemlerinde önemli bir rol üstlenmektedir (Koroteev, 2021). Google tarafından arama algoritmalarına entegre edilmesiyle birlikte, kullanıcıların sorgularına daha anlamlı ve bağlama uygun sonuçlar sunulması mümkün hâle gelmiştir. Bu gelişme, kullanıcı deneyiminin iyileşmesine ve arama sonuçlarının doğruluk ile fayda düzeyinin artmasına katkı sağlamıştır.

BERT, yalnızca bir model olmanın ötesinde, doğal dilin karmaşık yapısının çözümlenmesinde yeni bir yaklaşım ortaya koymuştur. Metin analizi ve çeviri uygulamalarında daha doğru sonuçlar elde edilmesini mümkün kılmasının temel nedeni, dildeki bağlamsal belirsizlikleri etkili biçimde modelleyebilmesidir. Örneğin, "banka" kelimesinin farklı anlamları, kullanım bağlamı dikkate alınarak ayırt edilebilmektedir. Bu başarının temelinde, Transformer mimarisine özgü dikkat (attention) mekanizması yer almaktadır. Söz konusu mekanizma, her kelimeyi cümledeki diğer kelimelerle olan ilişkileri çerçevesinde değerlendirerek bağlamsal temsiller üretmektedir. Bu nedenle BERT modeli, bir kelimenin hem öncesindeki hem de sonrasındaki bağlamsal ilişkileri eş zamanlı olarak analiz ederek metnin anlamını daha doğru biçimde yorumlayabilmektedir.



Şekil 1.3 BERT mimarisinin basitleştirilmiş görünümü (Vaswani vd., 2017)

Transformer mimarisinin kodlayıcı (encoder) ve çözücü (decoder) yapısı (Vaswani vd., 2017), Şekil 1.3'te sunulmaktadır. Bu görselde, BERT modelinin temelini oluşturan Transformer tabanlı kodlayıcı (encoder) yapısı, çift yönlü çalışma prensibi ve dikkat mekanizmasının işleyişi basitleştirilmiş bir biçimde gösterilmektedir.

Biyoinformatik ve yapay zekâ alanlarındaki gelişmeler, biyolojik verilerin analiz edilmesine yönelik güçlü yöntemlerin ortaya çıkmasını sağlamıştır. Son yıllarda özellikle DNA dizilerinin analizinde derin öğrenme ve doğal dil işleme (NLP) yaklaşımları ön plana çıkmaktadır. DNA metilasyonu, gen ekspresyonunun epigenetik düzeyde düzenlenmesinde önemli bir rol oynamakta olup hastalık tanısı, biyobelirteç keşfi ve genom yapısının anlaşılması gibi birçok kritik araştırma alanında yer almaktadır.

Geliştirilen derin öğrenme tabanlı modeller, DNA metilasyon bölgelerinin yüksek doğrulukla tahmin edilmesine olanak sağlamaktadır. Bu durum, deneysel yöntemlere olan bağımlılığı azaltmakta ve analiz süreçlerinin daha hızlı yürütülmesine katkı sunmaktadır (Yamamoto vd., 2024). Örneğin, birden fazla Transformer tabanlı dil modelinin birlikte kullanıldığı çalışmalar, 6mA, 4mC ve 5hmC olmak üzere üç farklı DNA metilasyon türünün başarılı biçimde tahmin edilebildiğini göstermektedir (Zeng vd., 2023). Bu bulgular, doğal dil işleme tekniklerinin biyolojik dizi analizlerinde de etkili bir şekilde uygulanabildiğini ortaya koymaktadır.

Biyolojik dizilerin analizinde klasik makine öğrenmesi yöntemlerinden derin öğrenme yaklaşımlarına geçiş, alan içerisinde önemli bir dönüşüm yaratmıştır. Özellikle DNA dizilerinin sınıflandırılması gibi karmaşık problemlerde, bu yeni yöntemlerin yüksek performans sağladığı görülmektedir. Son dönemde doğal dil işleme alanından uyarlanan yaklaşımlar da biyoinformatik çalışmalarına dâhil edilmiştir. DNA'nın bir dil yapısı gibi ele alınabilmesi, biyolojik verilerin modellenmesini önemli ölçüde kolaylaştırmıştır. Dört nükleotitten oluşan DNA dizileri ve bu dizilerde yer alan bağlamsal örüntüler, Transformer tabanlı modellerin biyolojik dizi analizlerine uygulanmasına olanak tanımıştır.

Bu yaklaşımdan hareketle geliştirilen Nucleic Transformer modeli, yalnızca yerel motifleri değil aynı zamanda uzun menzilli bağımlılıkları da birlikte öğrenebilen bir yapı sunmaktadır. Bu sayede DNA dizilerinin sınıflandırılmasında hem yüksek doğruluk elde edilmekte hem de yorumlanabilir bir model yapısı sağlanmaktadır (He vd., 2023). Araştırmacılar söz konusu modeli *E. coli* promotörlerinin ayırt edilmesi, insan ve fare genomlarında TATA ve non-TATA promotörlerinin sınıflandırılması,

viral genomların tanımlanması ve kromatin profilinin tahmin edilmesi gibi çeşitli biyoinformatik görevlerde değerlendirmiş ve tüm uygulamalarda etkili sonuçlar elde etmiştir (He vd., 2023).

Bu çalışmada, geleneksel yöntemlerin karşılaştığı zorluklar ve sınırlılıklar ele alınarak derin öğrenme modellerinin bu alandaki performansı incelenmekte ve söz konusu yaklaşımların etkinliğinin ortaya konulması amaçlanmaktadır. Özellikle DNA dizisi sınıflandırma probleminde BERT tabanlı bir modelin kullanılması, literatürdeki mevcut boşluğu doldurabilecek güncel bir yaklaşım sunmaktadır. Bu doğrultuda, klasik biyoinformatik yöntemlerin yetersiz kaldığı durumlarda derin öğrenme yöntemleriyle elde edilebilecek kazanımların gösterilmesi ve alana yenilikçi katkılar sağlanması hedeflenmektedir.

Bu çalışmanın temel araştırma sorusu, "BERT tabanlı bir modelin insan DNA dizilerinin sınıflandırılmasındaki performansı ne düzeydedir?" şeklinde belirlenmiştir. Bu doğrultuda ortaya konulan hipotez, "BERT tabanlı modellerin, geleneksel biyoinformatik yöntemlere kıyasla insan DNA dizilerinin sınıflandırılmasında daha yüksek doğruluk ve performans sağlayacağı" yönündedir.

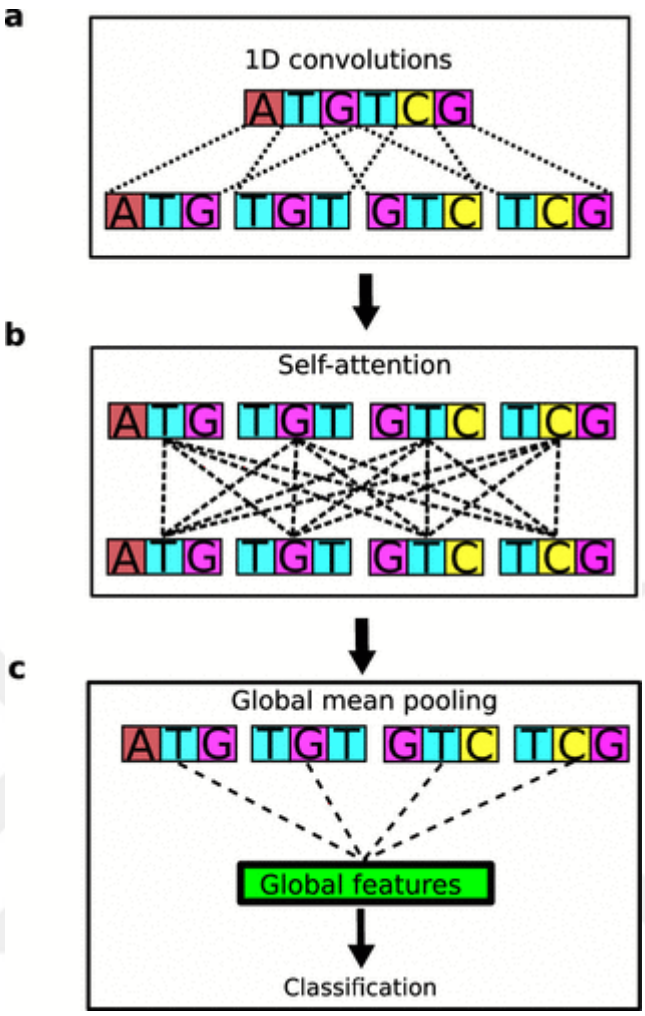
Bu hipotezin test edilmesi amacıyla BERT modeli DNA dizileri üzerinde uygulanacak ve elde edilen sonuçlar doğrultusunda modelin biyoinformatik alanındaki potansiyeli değerlendirilecektir. BERT modelinin DNA dizilerinde yer alan karmaşık örüntüleri yakalayabilme ve doğru sınıflandırma gerçekleştirebilme yeteneğinin, genomik araştırmalarda hem doğruluk hem de verimlilik açısından önemli katkılar sunabileceği öngörülmektedir.

Bu çalışma, DNA sınıflandırması gibi karmaşık bir problemi ele almakta ve biyoinformatik ile genetik alanlarında önemli bir teknolojik katkı sunmayı amaçlamaktadır. Önerilen model, yalnızca teorik bir yaklaşım ortaya koymakla sınırlı kalmayıp, uygulamada kullanılabilir bir çerçeve sunmayı hedeflemektedir. Bu yönüyle, gelecekte gerçekleştirilecek araştırmalar için yol gösterici bir temel oluşturma potansiyeline sahiptir. Genomik verilerin anlamlandırılması, günümüz biyoinformatik çalışmalarının temel gerekliliklerinden biridir. Özellikle virüslerin ve

genetik hastalıkların teşhisinde DNA dizilerinin doğru biçimde sınıflandırılması, tanı ve tedavi süreçlerinin hem hızlanmasına hem de daha hassas hâle gelmesine katkı sağlamaktadır (Sadad vd., 2023; Shirasawa ve Aguilar Cobos, 2024).

Geleneksel analiz yöntemlerinin bazı durumlarda yetersiz kalması, derin öğrenme tabanlı yaklaşımların bu alanda ön plana çıkmasına neden olmuştur. Son yıllarda özellikle Transformer mimarileri ve BERT modeli, doğal dil işleme alanındaki başarılarının DNA dizisi analizlerine uyarlanmasıyla biyoinformatik çalışmalarına yeni bir bakış açısı kazandırmıştır. Bu tez kapsamında, insan DNA dizilerinin sınıflandırılmasında BERT tabanlı yaklaşımların performansının ortaya konulması ve geliştirilen modelin tanı süreçlerindeki uygulanabilirliğinin değerlendirilmesi amaçlanmaktadır.

DNA dizileri, canlıların genetik kodunun temelini oluşturmakta ve genomik araştırmaların merkezinde yer almaktadır. Bu dizilerin doğru şekilde sınıflandırılması; gen fonksiyonlarının belirlenmesi, hastalık teşhisi ve genetik mühendisliği uygulamaları başta olmak üzere birçok alanda kritik bir rol oynamaktadır. Son yıllarda derin öğrenme, özellikle Transformer tabanlı modeller aracılığıyla DNA dizisi sınıflandırma problemlerinde önemli bir potansiyel ortaya koymuştur. BERT modeli ve biyolojik veri setlerine uyarlanmış türevleri, doğal dil işleme mimarilerini DNA analizine entegre ederek genetik verilerin yorumlanmasında yeni bir yaklaşım sunmuştur. Bu çalışmada, insan DNA dizilerinin sınıflandırılması amacıyla BERT tabanlı bir yöntem önerilmekte ve söz konusu yaklaşım mevcut yöntemlerle karşılaştırmalı olarak analiz edilmektedir. Çalışmanın temel amacı, genetik veri analizinde derin öğrenme tabanlı yöntemlerin gelecekteki uygulamalarına katkı sağlayacak sağlam bir temel oluşturmaktır.



Şekil 1.4 BERT tabanlı DNA sınıflandırma süreci

BERT tabanlı DNA dizisi sınıflandırma akış şeması (1D evrişim, öz-dikkat ve global havuzlama), Şekil 1.4'te sunulmaktadır. Bu görselde, DNA dizilerinin giriş aşamasından (ham diziler), ön işleme adımlarına (tokenizasyon ve embedding), verilerin BERT modeline aktarılmasına, sınıflandırma çıktısının elde edilmesine ve hastalık tanısı ile gen fonksiyonu belirleme gibi potansiyel uygulama alanlarına kadar uzanan süreç adım adım gösterilmektedir.

1.2. Literatürdeki Boşluklar ve Bu Tezin Katkıları

Bu tez ve çalışmada incelenen mevcut çalışmalar, DNA dizisi analizi alanında önemli ilerlemeler kaydetmiş olsa da, hala çeşitli eksiklikler ve geliştirilmesi gereken alanlar bulunmaktadır. Bu tez, bu boşlukları doldurarak DNA dizisi sınıflandırmasında daha

doğru, genellenebilir ve yorumlanabilir derin öğrenme modelleri geliştirmeyi amaçlamaktadır.

1.2.1. DNA'ya Özel Tokenizasyon Eksikliği ve Geliştirilmesi

Çoğu Transformer tabanlı çalışma, doğal dil işleme için geliştirilen tokenizasyon yöntemlerini örneğin k-mer ayrımını neredeyse hiç değişiklik yapmadan DNA dizilerine uygulamaktadır (Taylor vd., 2023). Oysa DNA dizileri sıradan metinler değildir; kendi içinde özgün biyolojik anlamlar, işlevsel motifler ve tekrarlayan yapılar barındırmaktadır. Mevcut k-mer tabanlı tokenizasyon yöntemleri, bu karmaşık bağlam ve hiyerarşiyi tam olarak yakalayamamaktadır. Bu çalışmada, klasik karakter tabanlı tokenizasyonun ötesine geçilmiştir. DNA'daki biyolojik motifleri ve kodon yapılarını daha iyi temsil eden bir k-mer tokenizasyon yöntemi uygulanmıştır. Bu sayede model, DNA dizilerini biyolojik açıdan daha anlamlı bir biçimde analiz edebilmektedir. Söz konusu yaklaşım, BERT'in bağlamsal öğrenme gücünü DNA'nın yerel motifleriyle birleştirmek adına önemli bir adım teşkil etmektedir.

1.2.2. Tam Genom Ölçeğinde Analiz ve Uzun Menzilli Etkileşimlerin Modellenmesi Sınırlılığı

Mevcut modeller genellikle kısa diziler veya belirli gen bölgeleri promotörler, bağlanma bölgeleri gibi üzerinde yoğunlaşmaktadır. Tam genom ölçeğinde, milyonlarca nükleotidi kapsayan uzun menzilli etkileşimleri (örneğin kromatin yapısı veya enhancer-promotör etkileşimleri) yakalayabilen ve hesaplama açısından verimli modellerin geliştirilmesi hâlen önemli bir güçlük arz etmektedir (Zhang vd., 2020). Geleneksel Transformer modellerinin dikkat (attention) mekanizmaları, dizi uzunluğu ile birlikte kuadratik olarak ölçeklendiğinden, çok uzun diziler için hesaplama yükünü artırmaktadır. Bu çalışmada, Longformer (Beltagy vd., 2020) veya Reformer gibi, daha uzun dizileri işleyebilen ve genomik bağlamı daha etkin şekilde kavrayabilecek BERT tabanlı modellerin uyarlanması veya yeni mimarilerin geliştirilmesi potansiyeli araştırılacaktır. Bu yaklaşım sayesinde yalnızca lokal motiflerin değil, tüm genomik bağlamın sınıflandırma üzerindeki etkisi incelenebilecektir.

1.2.3. Yorumlanabilirlik Sorunu ve Biyolojik İçgörülerin Elde Edilmesi

Derin öğrenme modelleri, karmaşık yapıları nedeniyle literatürde sıklıkla "kara kutu" olarak nitelendirilmektedir; bu durum, modelin karar verme süreçlerinin şeffaflığını kısıtlamaktadır. Biyolojik verilerin analizinde, modelin bir tahmini hangi DNA bölgelerine dayanarak yaptığı bilgisi, sonucun kendisi kadar kritiktir (Eraslan vd., 2019). Bu bağlamda, BERT tabanlı modellerin dikkat (attention) mekanizmalarının görselleştirilmesi ve özellik önem skorlarının belirlenmesi, biyolojik süreçlere dair yeni içgörüler sunma potansiyeline sahiptir (Ji vd., 2021). Çalışmada, LIME (Ribeiro vd., 2016) ve SHAP (Lundberg ve Lee, 2017) gibi açıklanabilir yapay zeka (XAI) yöntemlerinin uyarlanmasıyla, modelin biyolojik olarak anlamlı motifleri ve mekanizmaları ne ölçüde ayırt edebildiği detaylandırılacaktır.

Buradaki amaç, mevcut eksiklikleri gidererek DNA dizisi sınıflandırmasında daha doğru, genellenebilir ve yorumlanabilir derin öğrenme modelleri geliştirmektir. Özellikle DNA'ya özel olarak tasarlanmış bir BERT modelinin, farklı DNA dizisi türlerini ne derece yüksek doğrulukla ayırt edebildiği değerlendirilecektir. Bu çalışma, yalnızca teorik bir katkı sunmakla kalmayıp, biyolojik araştırmalarda ve klinik uygulamalarda DNA dizisi analizine yönelik pratik ve yenilikçi bir araç geliştirmeyi de hedeflemektedir.

1.3. Sınırlamalar

Çalışmanın bazı sınırlamaları şunlardır:

1. Veri Seti Boyutu: Çalışmada kullanılan 4.380 DNA dizisi, derin öğrenme modelleri açısından görece küçük bir veri seti olarak değerlendirilmektedir. Daha geniş veri setleri kullanılması durumunda, modelin performansının önemli ölçüde artırılabileceği öngörülmektedir.

2. Sınıf Dengesizliği: Bazı sınıflarda örnek sayısı diğer sınıflara kıyasla önemli ölçüde azdır. Bu durum, modelin söz konusu sınıflarda yeterli performans göstermesini engellemiştir. Gelecekte, veri artırma teknikleri veya ağırlıklı kayıp

fonksiyonları gibi yöntemler kullanılarak bu dengesizliğin giderilmesi mümkün olabilir.

3. Tokenizasyon: DNA dizilerini karakter bazında tokenize etmek, bazı bağlamsal bilgilerin kaybolmasına yol açmaktadır. Bunun yerine, k-mer tabanlı tokenizasyon gibi alternatif yöntemler değerlendirilebilir (Ji vd., 2021).

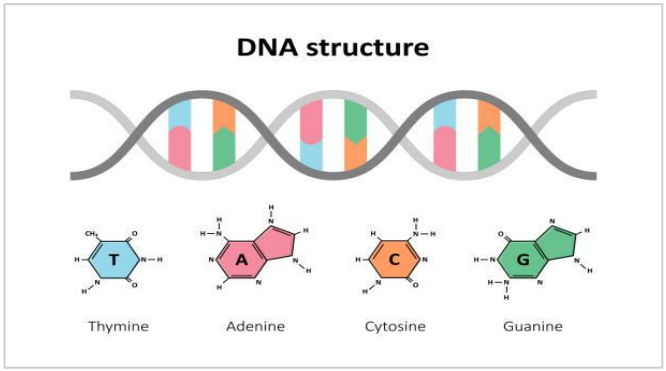
2. LİTERATÜR ARAŞTIRMASI

DNA dizilerinin sınıflandırılması, biyoinformatik alanında uzun süredir üzerinde çalışılan temel konulardan biridir. Bu bölümde öncelikle DNA sınıflandırmasında kullanılan geleneksel yöntemler ele alınmakta, ardından derin öğrenme modellerine ve özellikle BERT gibi Transformer tabanlı yaklaşımlara odaklanılmaktadır. Bu kapsamda, söz konusu modellerin bu alanda nasıl kullanıldığı ayrıntılı biçimde incelenmektedir. Ayrıca mevcut yöntemlerin güçlü ve sınırlı yönleri değerlendirilerek, bu tezin literatürdeki boşluğu nasıl doldurduğu tartışılmaktadır.

2.1. DNA (Deoksiribonükleik Asit)

Deoksiribonükleik asit (DNA), canlıların genetik bilgisini taşıyan temel moleküldür. Hemen hemen tüm canlı organizmalarda ve birçok virüste genetik bilginin temel depolama birimi olarak görev yapmaktadır. DNA'nın keşfi ve yapısının aydınlatılması, modern biyolojide önemli bir dönüm noktası oluşturmuş; genetik, moleküler biyoloji ve biyoteknoloji başta olmak üzere birçok alanda önemli gelişmelerin önünü açmıştır (Lesk, 2013; Smith ve Brown, 2019). DNA, bir organizmanın gelişimi, büyümesi, çoğalması ve hayati işlevlerini sürdürebilmesi için gerekli olan genetik talimatları içermektedir. Bu yönüyle DNA, yaşamın biyolojik planını oluşturan temel yapı olarak değerlendirilmektedir.

DNA'nın moleküler yapısı, 1953 yılında James Watson ve Francis Crick tarafından önerilen modele dayanmaktadır. Buna göre DNA, birbirine antiparalel konumlanmış iki polinükleotid zincirinden oluşan çift sarmal bir yapıya sahiptir. Her nükleotid üç temel bileşenden meydana gelmektedir: bir deoksiriboz şekeri, bir fosfat grubu ve dört azotlu bazdan biri olan adenin (A), timin (T), sitozin (C) veya guanin (G). Şeker ve fosfat grupları sarmal yapının dış kısmını oluşturarak yapısal bir omurga görevi görürken, azotlu bazlar içe doğru uzanarak karşı zincirdeki tamamlayıcı bazlarla eşleşmektedir. Bu eşleşmede adenin timinle (A–T) iki hidrojen bağı, guanin ise sitozinle (G–C) üç hidrojen bağı kurmaktadır (Watson ve Crick, 1953). Bu özgül baz eşleşmesi, çift sarmal yapının kararlılığını sağlamakta ve DNA'nın replikasyon ile transkripsiyon süreçleri sırasında genetik bilginin doğru biçimde aktarılmasına olanak tanımaktadır (Cabral-de-Mello vd., 2024).



Şekil 2.1 DNA'nın Çift Sarmal Yapısı ve Baz Bileşenleri(iStock, n.d.).

Şekil 2.1'de DNA'nın çift sarmal yapısı gösterilmektedir. Bu yapıda iki adet şeker-fosfat omurgası birbirine antiparalel şekilde uzanmakta ve aralarında karşılıklı nükleotid baz eşleşmeleri bulunmaktadır. Baz eşleşmeleri adenin–timin (A–T) ve guanin–sitozin (G–C) çiftleri arasında gerçekleşmektedir. Görselin alt kısmında ise timin, adenin, sitozin ve guanin bazlarının kimyasal yapı formülleri verilerek bu bazların moleküler farklılıkları ve birbirleriyle kurdukları tamamlayıcı eşleşme düzeni görselleştirilmektedir (iStock, n.d.).

DNA'nın temel işlevi, genetik bilginin uzun süreli olarak depolanmasıdır. Bu bilgi, azotlu bazların doğrusal dizilimi içinde kodlanmıştır ve dört harfli bir alfabe (A, T, C ve G) gibi düşünülebilir. Belirli baz dizilimleri, proteinleri veya işlevsel RNA moleküllerini kodlayan genleri oluşturur ve bu yapı canlı organizmaların genetik temelini meydana getirir (Alberts vd., 2015).

Genetik bilginin DNA'dan işlevsel ürünlere aktarımı, moleküler biyolojide merkezi dogma (central dogma) olarak bilinen süreç ile açıklanır. Bu modele göre genetik bilgi DNA'dan RNA'ya ve RNA'dan proteine doğru aktarılır (Crick, 1970). İlk aşamada DNA'nın belirli bir bölgesi, tamamlayıcı bir haberci RNA (mRNA) molekülüne kopyalanır; bu süreç transkripsiyon olarak adlandırılır. Daha sonra ribozomlar mRNA dizisini okuyarak belirli bir proteinin sentezlenmesini sağlar; bu aşama ise translasyon olarak adlandırılır. Bu iki aşamadan oluşan gen ifadesi mekanizması, hücresel faaliyetlerin düzenlenmesinde ve organizmanın fenotipik özelliklerinin ortaya çıkmasında temel bir rol oynar (Yang vd., 2024).

DNA sadece protein sentezinden sorumlu değil, başka önemli işlere de bakmaktadır.

- Genetik Kalıtım: DNA, kalıtımın temel unsurudur. Eşeyli veya eşeysiz üreme süreçlerinde DNA'daki genetik bilgi ebeveynlerden yavrulara aktarılır. DNA replikasyonu sayesinde her yeni hücre, genetik materyalin birebir kopyasını alır. Bu sayede kalıtsal özellikler nesilden nesile aktarılmaktadır.

- Varyasyon ve Evrim: DNA replikasyonu genellikle doğru bir şekilde gerçekleşir; ancak zaman zaman hatalar meydana gelir ve bu durum mutasyon olarak adlandırılır. Mutasyonlar ile genetik rekombinasyon, bir popülasyondaki genetik çeşitliliğin başlıca kaynaklarını oluşturur. Bu çeşitlilik, doğal seçilimin etkisiyle evrimi şekillendirir ve canlıların çevresel koşullara uyum sağlamasına olanak tanır (Cabral-de-Mello vd., 2024).

- Gen İfadesinin Düzenlenmesi: DNA'nın tamamı protein kodlamamaktadır. Özellikle ökaryotlarda, genomun büyük bir kısmı genlerin ne zaman ve nerede ifade edileceğini kontrol eden dizilerden oluşur. Promoter, enhancer ve silencer olarak adlandırılan bu bölgeler, DNA'ya bağlanan proteinler ve transkripsiyon faktörleri ile etkileşime girerek gen ifadesini, gelişim süreçlerine veya çevresel değişimlere göre ince bir şekilde düzenler (Öncül, 2022).

DNA'nın işleyişini çözmemiz, teknolojide tamamen yeni bir dönemin kapılarını açtı. Genetik mühendisliği sayesinde artık bir canlının genomu doğrudan değiştirilebilmektedir. CRISPR-Cas9 gibi sistemler, DNA dizilerini neredeyse kusursuz ve son derece etkili bir biçimde düzenlemeyi mümkün kılmaktadır (Smith ve Brown, 2019). Bu gelişmelerin etkisi oldukça geniş kapsamlıdır; tıpta gen terapisi ve kişiye özel tedavi yöntemleri ortaya çıkmış, tarımda genetiği değiştirilmiş bitkiler hayatımıza girmiştir. Ayrıca çevre bilimlerinde biyoteknolojik arıtma gibi alanlarda da köklü değişiklikler gözlemlenmiştir (Chen ve Zhang, 2020).

Artık araştırmacılar DNA'yı yalnızca tek genler üzerinden incelememekte, tüm genomu kapsamlı bir şekilde ele almaktadır. Genomik, metagenomik ve filogenetik gibi alanlarda çalışmalar, tamamen DNA dizileme, analiz ve yorumlama becerisine dayanmaktadır. Genomik, tüm genomları incelerken; metagenomik, çevreden alınan örneklerdeki mikrobiyal toplulukların genetik yapısını değerlendirir; filogenetik ise canlılar arasındaki evrimsel ilişkileri araştırır (Minamoto, 2022). Ayrıca epigenetik

alanında, DNA dizisi değişmeden genlerin farklı şekilde nasıl çalıştığı ve bu değişikliklerin nasıl kalıtıldığı incelenmektedir. Özellikle DNA metilasyonu gibi süreçlerin, genlerin düzenlenmesinde kritik bir rol oynadığı ve birçok hastalıkla ilişkili olduğu bilinmektedir (Yamamoto vd., 2024).

Sonuç olarak, DNA yalnızca hareketsiz bir bilgi deposu değildir; yaşamın merkezinde, sürekli aktif bir molekül olarak işlev görür. Çift sarmal yapısı sayesinde genetik bilgiler güvenli bir şekilde saklanır ve doğru biçimde kopyalanır. DNA dizilerindeki çeşitlilik, canlılardaki şaşırtıcı farklılığın temel anahtarını oluşturur. DNA üzerine yapılan araştırmalar, biyoloji ve tıp alanında sürekli yeni kapılar açmakta; hem yaşamın işleyişini daha iyi anlamamıza yardımcı olmakta, hem de insan sağlığının iyileştirilmesine katkı sağlamaktadır.

2.2. DNA Sınıflandırmasında Geleneksel Yöntemler ve Sınırlamaları

DNA sınıflandırmasında, araştırmacılar uzun süredir istatistiksel yöntemler ve makine öğrenmesi algoritmalarından yararlanmaktadır. Bu yöntemler çoğunlukla dizilerden elde edilen ve manuel olarak seçilen özelliklere dayanmaktadır. Belirli görevlerde zaman zaman iyi sonuçlar elde edilse de, karmaşık ve uzun dizilerdeki ince yapıları yakalamada sınırlı başarı göstermektedirler. En yaygın olarak kullanılan yöntemler ise şunlardır:

2.2.1 Destek Vektör Makineleri (SVM) ve Karar Ağaçları

Destek Vektör Makineleri (SVM'ler), DNA bağlayıcı proteinlerin tespiti, promotör bölgelerin tahmini ve genlerin sınıflandırılması gibi görevlerde sıklıkla tercih edilmektedir (Liu ve Wang, 2015). Bu yöntemler, genellikle DNA dizilerinden elde edilen pseudo amino asit kompozisyonu (PseAAC), k-mer frekansları veya fizikokimyasal özellikler gibi manuel olarak çıkarılmış özellikler üzerinden bir sınıflandırma düzlemi oluşturur. Örneğin, SVM tabanlı bir modelle DNA bağlayıcı proteinler %82 doğrulukla ayrılabilmiştir (Qu ve Zou, 2019). Benzer şekilde, SVM ile genetik algoritmaların birleştirilmesi de DNA bağlayıcı proteinlerin tespitinde başarılı sonuçlar vermiştir. Karar ağaçları ise "eğer–ise" kuralları ile DNA dizilerini sınıflandırmakta ve özellikle anlaşılabilir olmalarıyla ön plana çıkmaktadır. Ancak bu

yöntemlerin temel bir sınırlılığı, başarılarının tamamen elde edilen özelliklerin kalitesine bağlı olmasıdır. Yani, DNA dizilerinin karmaşık ve katmanlı yapısını tam olarak yansıtamamaktadırlar. Ayrıca büyük veri setlerinde veya özellik sayısının artması durumunda hesaplama yükü önemli ölçüde yükselmektedir.

2.2.1. k-mer Tabanlı Yaklaşımlar

DNA dizilerinin frekans dağılımını incelemenin en yaygın yöntemlerinden biri, k-mer tabanlı yaklaşımlardır. Genellikle 3 ila 6 nükleotitten oluşan kısa alt diziler kullanılmaktadır. Bu yöntemler, gen ifadesi analizi, tür tanımlama, metagenomik çalışmalar ve türe özgü belirlemelerde sıklıkla tercih edilmektedir (Lebatteux vd., 2019). K-mer frekansları, bir dizinin yapısal özelliklerini özetlemenin basit ancak etkili bir yolunu sunmaktadır. Örneğin, viral genomların sınıflandırılmasında k-mer özelliklerini kullanan makine öğrenimi modelleri hızlı ve doğru sonuçlar elde etmektedir (Zhang vd., 2023). Ayrıca DNA dizileri, k-mer özellikleriyle grafiklere dönüştürülerek grafik sinir ağlarıyla sınıflandırıldığında %88 doğruluk oranı sağlanmıştır. Ancak k-mer tabanlı yöntemlerin bazı sınırlılıkları bulunmaktadır. Bu yöntemler, dizideki uzun menzilli bağımlılıkları veya motiflerin uzamsal ilişkilerini doğrudan tespit edememektedir. Ayrıca, k-mer uzunluğu arttıkça özellik uzayı hızla genişlemekte, bu da "boyut laneti" olarak adlandırılan soruna yol açmakta ve verilerin hızla seyrekleşmesine neden olmaktadır.

2.2.2. Saklı Markov Modelleri (HMM)

Saklı Markov Modelleri (HMM'ler), DNA'daki protein kodlayan ve kodlamayan bölgelerin belirlenmesi (gen bulma), promotör bölgelerinin tahmini ve CpG adalarının tespitinde kullanılmaktadır. HMM'ler, DNA dizilerindeki ardışık bağımlılıkları ve biyolojik süreçlerin altında yatan gizli durumları modelleyebilme yeteneği sayesinde, gen bulma ve gen ifadesi analizi gibi alanlarda önemli uygulamalara sahiptir (Durbin vd., 1998). Bu modeller, her bir nükleotidin bir durumdan diğerine geçiş olasılıklarını ve belirli bir durumda bir nükleotidin gözlenme olasılığını modelleyerek dizilerin olasılıksal bir temsilini sunar. Ancak HMM'ler, genellikle yalnızca lokal dizilim örüntülerini yakalayabilmekte olup, karmaşık epigenetik regülasyon mekanizmalarını veya uzun menzilli etkileşimleri modellemede yetersiz kalmaktadır. Modelin

karmaşıklığı, gerçek biyolojik süreçlerin tamamını kapsamakta sınırlı kalabilir ve özellikle büyük ve çeşitli genomlar için genelleme yeteneği sınırlıdır.

2.2.3. Geleneksel Yöntemlerin Başlıca Sınırlamaları

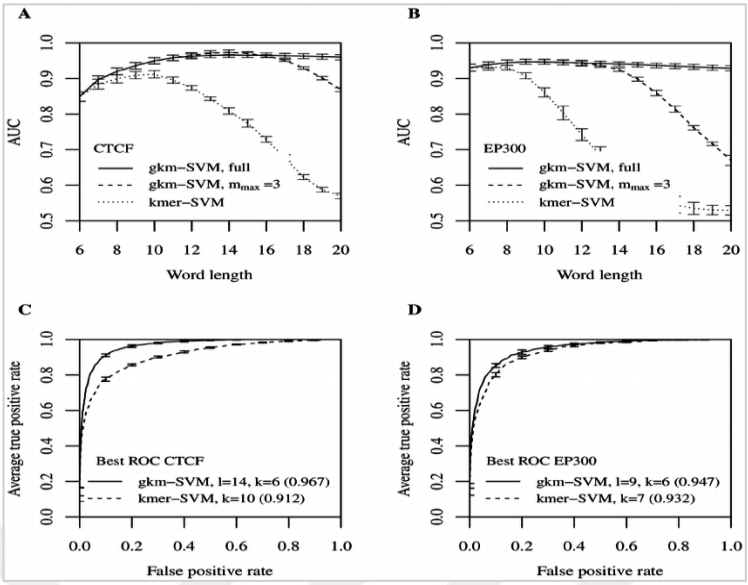
1. Manuel Özellik Çıkarımı Gerektirmesi: Geleneksel makine öğrenmesi algoritmaları, DNA dizilerinden doğru özellikleri elde etmek için kapsamlı biyoloji bilgisine ihtiyaç duyar. Bu süreç hem zaman alıcıdır hem de hata yapma olasılığı yüksektir. Veri seti büyüdükçe işler daha da karmaşık hâle gelir ve verim düşer. Ayrıca önemli veya henüz keşfedilmemiş özelliklerin gözden kaçması kolaydır; bu durum modelin performansını olumsuz etkiler ve süreci daha maliyetli hâle getirir.

2. Uzun Menzilli Bağımlılıkları Modelleyememe: Promotör–enhancer etkileşimleri veya alternatif ekleme bölgeleri gibi DNA'daki uzun mesafeli ilişkiler, klasik yöntemler tarafından genellikle yeterince yakalanamaz. Bu durum, gen regülasyonu gibi karmaşık biyolojik süreçlerin tam olarak anlaşılmasını ve doğru tahmin edilmesini güçleştirir. Özellikle fonksiyonel DNA bölgelerinin tespiti ve hastalıkların ortaya çıkış mekanizmalarının anlaşılması söz konusu olduğunda, bu ciddi bir dezavantaj yaratır.

3. Düşük Genelleme Performansı: Klasik modeller, eğitildikleri veri setine aşırı uyum sağlayabilir. Bu nedenle, başka organizma türleri veya yeni veri setleri ile karşılaştıklarında performansları düşer. Bir tür için iyi çalışan bir modeli farklı bir türde veya farklı deneysel koşullarda kullanmak çoğu zaman mümkün değildir. Bu da her yeni durumda veya farklı türlerde modelin yeniden kurulmasını ve özelliklerin tekrar tasarlanmasını gerektirir.

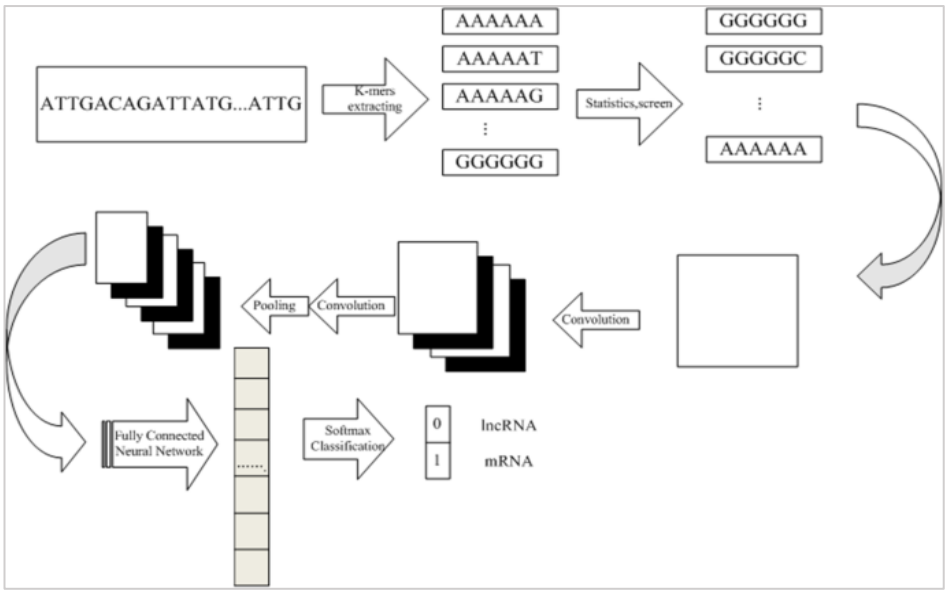
4. Hesaplama Verimsizliği: Büyük genomik veri setleriyle çalışırken, k-mer tabanlı yöntemler yüksek boyutlu özellik uzayları ile, HMM'ler ise karmaşık olasılık hesapları ile ilgilenir. Bu durum ciddi ölçüde bilgisayar gücü ve zaman gerektirir. Özellikle genom çapında analiz yapılması gerektiğinde, bu süreç neredeyse uygulanamaz hâle gelir.

DNA Sınıflandırmasında Kullanılan Geleneksel Yöntemlerin Şematik Gösterimi:



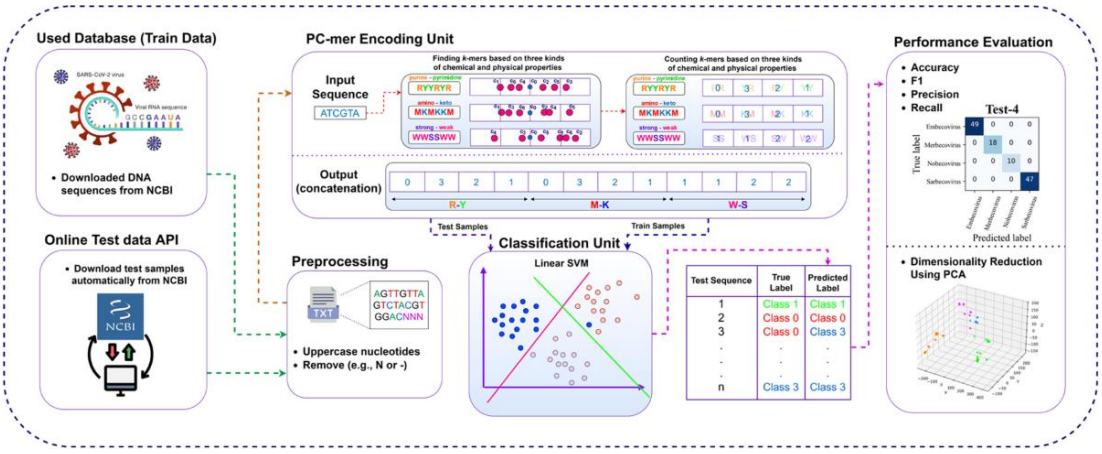
Şekil 2.2 gkm-SVM vs. k-mer-SVM performans grafiği (Ghandi vd., 2014).

Şekil 2.2 görselde, geleneksel k-mer SVM ile gkm-SVM'in AUC değerlerini karşılaştıran istatistiksel sonuç görüntüsünü içerir. Özellikle CTCF ve EP300 bağlanma bölgelerinde gkm-SVM'in üstünlüğü vurgulanır(Ghandi vd., 2014).



Şekil 2.3 k-mer histogram + CNN sınıflandırma akış diyagramı(Zhang vd., 2019)

Şekil 2.3 görselde, DNA dizisinden k-mer frekanslarının çıkarılması, görüntü-matris haline dönüştürülüp CNN'e beslenmesini modellemektedir(Zhang vd., 2019).



Şekil 2.4 PC-mer encodlama ve sınıflandırma süreci (Akbari Rokn Abadi vd., 2023).

Şekil 2.4 görselde, PC-mer yönteminin pipeline'ını; encode işlemini ve ardından sınıflandırma adımlarını özetlemektedir(Akbari Rokn Abadi vd., 2023).

2.3. Derin Öğrenme Tabanlı DNA Sınıflandırma Yöntemleri

Derin öğrenme modelleri, ham DNA dizilerinden kendileri özellik çıkarabildikleri için son yıllarda biyoinformatikte iyice yaygınlaşmaktadır. Bu modeller, karmaşık ve katmanlı özellikleri öğrenerek klasik yöntemlerin önüne geçiyor. Temelde üç ana model tipi var.

2.3.1. Evrişimli Sinir Ağları (CNN)

Evrişimli Sinir Ağları (CNN'ler), DNA'daki korunmuş motifleri tespit etmede oldukça etkili bir performans sergiler. Örneğin, TATA kutuları veya transkripsiyon faktörü bağlanma bölgeleri gibi motifleri kolaylıkla tanıyabilir. Aslında, CNN'lerin görüntülerdeki küçük desenleri yakalama yeteneği, DNA dizilerindeki yerel motifleri bulmaya uyarlanmıştır. DNA dizileri, tek boyutlu bir görüntü olarak ele alınır ve evrişim filtreleriyle taranır. Bu filtreler, dizideki belirli kısa nükleotid dizilerini öğrenir ve motiflerin varlığını saptar. DeepBind (Alipanahi vd., 2015) bu yaklaşımın başarılı bir örneğidir; RNA ve DNA bağlayıcı proteinlerin hedef bölgelerini tahmin etmek için CNN kullanmış ve klasik yöntemlere kıyasla %89 doğrulukla yüksek başarı elde etmiştir. Benzer şekilde, DeepSEA (Zhou ve Troyanskaya, 2015) modeli de CNN'lerin gücünü göstermektedir. DeepSEA, kromatin erişilebilirliği, histon modifikasyonları ve transkripsiyon faktörü bağlanma bölgeleri gibi epigenetik özellikleri genom

genelinde tahmin edebilmekte, toplamda 919 farklı epigenetik özelliği öngörerek genetik varyantların bu özellikler üzerindeki potansiyel etkilerini değerlendirebilmektedir. Özetle, CNN'ler özellikle motif tespiti ve kısa dizilerdeki yerel etkileşimlerin belirlenmesinde oldukça güçlüdür.

2.3.2. Tekrarlayan Sinir Ağları (RNN/LSTM)

Tekrarlayan Sinir Ağları (RNN'ler), sıralı verilerdeki ardışık bağımlılıkları anlamak amacıyla geliştirilmiştir. Ancak klasik RNN'ler, uzun diziler söz konusu olduğunda yeterli performans göstermez; zira "uzun menzilli bağımlılık" sorunuyla karşılaşırlar ve gradyanlar ya kaybolur ya da kontrol edilemez hâle gelir. Bu sorunun üstesinden gelmek için LSTM ve GRU gibi daha gelişmiş RNN türleri geliştirilmiştir. LSTM ağları, özellikle DNA'daki uzun dizilimlerin bağımlılıklarını modellemede oldukça etkilidir. Gen ifadesi analizi, splicing bölgelerinin belirlenmesi veya düzenleyici elementlerin tahmini gibi alanlarda kullanılmaktadır. Bu ağlar, uzun dizilerdeki karmaşık ilişkileri yakalamada başarılıdır. Örneğin, DanQ modeli (Quang ve Xie, 2016), CNN ve LSTM katmanlarını bir araya getirerek DNA'nın fonksiyonel anotasyonunda %91 doğruluk elde etmiştir. CNN motifleri tanımakta ustayken, LSTM dizisel bağımlılıkları öğrenir; ikisinin birleşimi, modelin çok daha kapsamlı bir şekilde öğrenmesini sağlar. RNN ve LSTM'ler, özellikle RNA'nın ikincil yapısını tahmin etmek veya genetik kodlamayı analiz etmek gibi sıralı bilgi akışının kritik olduğu durumlarda umut verici sonuçlar sunmuştur.

2.3.3. Hibrit Modeller (CNN-BiLSTM, CNN-Attention)

Son dönemde yapılan araştırmalar, farklı derin öğrenme mimarilerini bir araya getiren hibrit modellerin, DNA dizisi analizinde daha etkili ve kapsamlı çözümler sunduğunu ortaya koymaktadır. Bu hibrit yapılar sayesinde hem lokal hem de global özellikler eş zamanlı olarak öğrenilebilmektedir.

1. CNN-BiLSTM: Bu mimaride, CNN'in lokal motifleri yakalama yeteneği ile BiLSTM'in dizideki uzun menzilli bağımlılıkları iki yönlü olarak öğrenme gücü birleştirilmiştir (Wang vd., 2023). Araştırmacılar, bu hibrit modeli viral genom sınıflandırmasında denemiş ve yüksek bir başarı elde etmişlerdir. Ayrıca, DNA

dizisi sınıflandırmasında hem CNN'i hem de hibrit yaklaşımları karşılaştıran çalışmalar, hibrit modellerin sağladığı üstün performansı açıkça göstermektedir (Gunasekaran vd., 2021).

2. CNN-Attention: Dikkat mekanizmaları, modelin dizinin farklı bölümlerine farklı derecelerde önem atamasını sağlar. Böylece model, hangi kısımlardan bilgi alması gerektiğini kendi kendine öğrenebilmektedir; bu durum özellikle uzun diziler için büyük bir avantaj sunmaktadır (Chen vd., 2023).

Örneğin, DeepAttentioDNA modeli, dikkat mekanizmalarını kullanarak önemli nükleotid bölgelerini vurgulamış ve DNA dizisi sınıflandırmasında %92 doğruluk elde etmiştir. Dikkat mekanizmaları, modelin tahminlerini yorumlanabilir hâle getirme potansiyeli sunar; çünkü hangi bölgelerin modele en çok katkıda bulunduğunu gösterir. Bu tür modeller, karmaşık gen regülasyon ağlarının veya transkripsiyon faktörü bağlanma bölgelerinin daha hassas şekilde tanımlanmasında etkili bir şekilde kullanılmaktadır (Chen vd., 2023).

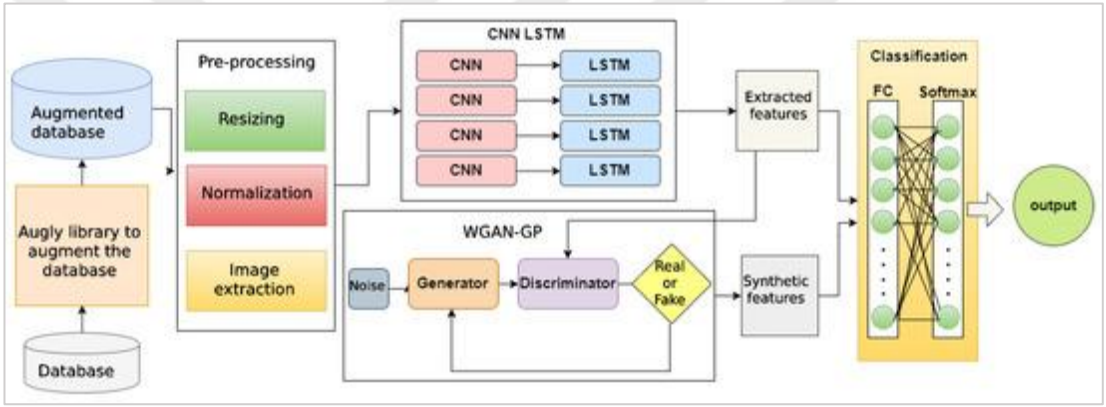
2.3.4. Derin Öğrenmenin Sınırlamaları

1. Büyük Veri İhtiyacı: Derin öğrenme modelleri genellikle büyük miktarda etiketli veriye ihtiyaç duymaktadır. Ancak biyolojik verilerde bu her zaman mümkün olmamaktadır. Özellikle nadir genetik varyantlar veya belirli hastalıklarla ilişkili DNA dizileri söz konusu olduğunda yeterli veri bulmak neredeyse imkânsız hale gelmektedir. Eğitim verisi az olduğunda, modelin genelleme yeteneği düşmekte ve aşırı öğrenme (overfitting) riski artmaktadır (Smith ve Brown, 2023). Bu tür durumlarda veri artırma (data augmentation) veya transfer öğrenme gibi yöntemlerin uygulanması neredeyse zorunlu hâle gelmektedir.

2. Uzun Dizilerde Hesaplama Zorluğu: Tam genom gibi çok uzun DNA dizileri üzerinde çalışırken, derin öğrenme modelleri; RNN'ler ve Transformer'lar dâhil, yüksek hesaplama gücü ve bellek talep etmektedir (Zhang vd., 2020). Bu durum, çalışmayı zorlaştırmakta ve güçlü donanım gereksinimini (GPU, TPU gibi) artırmaktadır. Longformer gibi daha verimli Transformer mimarileri bu yükü bir

ölçüde hafifletmeye çalışsa da, tam genom boyutunda hâlâ önemli hesaplama zorlukları mevcuttur (Beltagy vd., 2020).

3. Yorumlanabilirlik Eksikliği (“Kara Kutu” Sorunu): Derin öğrenme modelleri genellikle “kara kutu” gibi çalışmaktadır; yani, modelin bir tahmini neden yaptığı açıkça görülememektedir. Oysa biyolojik araştırmalarda, modelin hangi DNA bölgelerine odaklandığını veya hangi motifleri yakaladığını bilmek kritik öneme sahiptir. Bu bilgi, hem yeni içgörüler geliştirmek hem de yeni hipotezler üretmek için gereklidir. Yorumlanabilirliğin eksikliği, özellikle klinik uygulamalar ve ilaç geliştirme süreçlerinde modelin güvenilirliğini ve kabulünü ciddi şekilde sınırlamaktadır.



Şekil 2.5 Genel CNN-LSTM mimari şeması(Gulrajani vd., 2021).

Şekil 2.5 görselde, aerial video analizi için geliştirilen bir modelin mimari yapısı, CNN ile öznitelik çıkarımı + LSTM ile zaman boyutu modellendiği genel akış DNA sınıflandırma modellerine çok benzer yapıdadır. CNN (evrişim katmanları ve havuzlama), RNN/LSTM (tekrarlayan hücreler ve zaman adımları) ve hibrit modellerin (CNN-BiLSTM, CNN-Attention) temel yapılarını ve DNA dizisi verilerini nasıl işlediklerini şematik olarak göstermektedir. Giriş olarak DNA dizisi, çıktı olarak sınıflandırma sonucu gösterilmelidir.

2.4. Dönüştürücü (Transformer) Tabanlı Modeller ve BERT'in Uygulanması

Doğal dil işleme (NLP) alanındaki çığır açan başarısı nedeniyle Transformer mimarisi (Vaswani vd., 2017), DNA dizilerinin modellenmesinde de kullanılmaya başlanmıştır. Transformer modelleri, özellikle BERT (Bidirectional Encoder Representations from

Transformers), dil modellemede çift yönlü bağlam anlama yetenekleriyle öne çıkar. Bu özellik, bir kelimenin veya nükleotidin hem önceki hem de sonraki dizilim bağlamına göre anlamlandırılmasını sağlar. Bu durum, DNA dizilerindeki karmaşık ilişkileri ve uzun menzilli bağımlılıkları yakalamak için ideal bir yaklaşım sunmaktadır; çünkü DNA dizileri de, tıpkı insan dilleri gibi, belirli kurallara ve bağlama dayalı bir yapıya sahiptir. BERT, maskelenmiş dil modelleme ve sonraki cümle tahmini gibi ön eğitim görevleriyle büyük veri setleri üzerinde eğitilmekte olup, bu sayede downstream görevler (sınıflandırma, dizi etiketleme vb.) için hazır hâle gelmektedir.

BERT'in genomikteki uygulamaları:

- DNA-BERT: DNA-BERT, insan genomundaki DNA metilasyon bölgelerini (özellikle 5-metilsitozin, 5mC) tahmin etmek için özelleştirilmiş bir BERT modelidir. Geleneksel yöntemlere kıyasla %18 daha yüksek AUC (Alan Altında Eğri) değeri (0.94) elde ederek önemli bir performans artışı sağlamıştır. Bu model, DNA dizilerini dil benzeri bir yapıya dönüştürmek için 6-mer (hegzamer) tokenizasyonu kullanmıştır; yani DNA dizileri 6 nükleotitlik alt dizilere ayrılmış ve bunlar kelime (token) gibi işlenmiştir. DNA-BERT, özellikle epigenetik işaretlerin, yani DNA metilasyonunun doğru bir şekilde belirlenmesinde önemli başarılar kaydetmiştir. Bu, gen ekspresyonu ve hastalık gelişimi için kritik öneme sahip olan metilasyon kalıplarının anlaşılmasına büyük katkı sağlamıştır(Ji vd., 2021).

- Nucleic Transformer: Nucleic Transformer, DNA'daki hem kısa hem de uzun menzilli bağımlılıkları aynı anda modelleyebilen Transformer tabanlı bir modeldir. RNA ve DNA dizisi sınıflandırma görevlerinde başarılı bir şekilde uygulanmıştır. Model, E. coli promotör sınıflandırmasında %96 doğruluk sağlayarak dikkat çekici bir performans sergilemiştir. Ayrıca, modelin yorumlanabilirliği için öz-dikkat haritaları (self-attention maps) sunar. Bu haritalar, modelin bir tahminde bulunurken DNA dizisinin hangi kısımlarına odaklandığını görselleştirmeye olanak tanır. Bu, modelin DNA dizileri arasındaki karmaşık ilişkileri anlamada yeni bir kapı aralamış ve biyologlara modelin kararlarını anlamak için değerli içgörüler sunmuştur(He vd., 2023).

- MuLan-Methyl: MuLan-Methyl, BERT, DistilBERT ve ALBERT gibi çoklu Transformer modellerini entegre eden karmaşık bir mimaridir. Bu model, sadece bir metilasyon tipini değil, aynı anda 6mA (N6-metiladenozin), 4mC (N4-metilsitozin) ve 5hmC (5-hidroksimetilsitozin) gibi farklı metilasyon tiplerini tahmin edebilir. Model, taksonomik bilgileri (yani, organizmanın türü gibi ek bağlamsal bilgileri) de işleyerek tahmin doğruluğunu artırmış ve AUC'yi 0.97'ye çıkarmıştır. Bu, modelin farklı metilasyon türlerini yüksek hassasiyetle ayırt etme yeteneğini ve çoklu görev öğrenme potansiyelini göstermektedir. Bu tür modeller, epigenetik araştırmalar için güçlü araçlar sunarak gen regülasyonunun daha kapsamlı anlaşılmasına olanak tanır(Zeng vd., 2023).

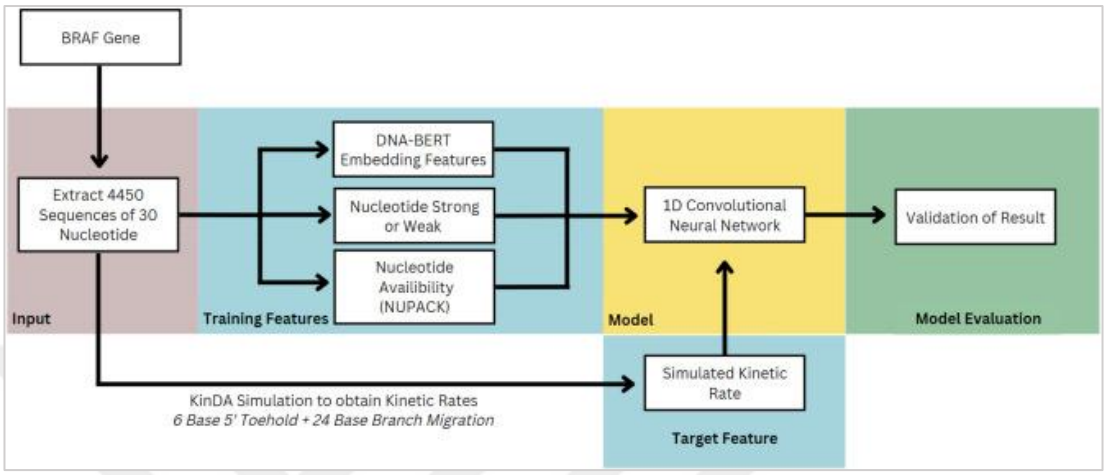
BERT'in Avantajları ise şunlardır:

- Çift Yönlü Bağlam Anlayışı: BERT, Transformer'ın çift yönlü kodlayıcı yapısını kullanmaktadır. Yani bir nükleotidi veya k-mer'i işlerken hem öncesine hem de sonrasına aynı anda bakmaktadır. Klasik, tek yönlü dil modelleri, örneğin RNN'ler, yalnızca önceki bağlamı dikkate alır. Ancak BERT, dizideki her bir elementi değerlendirirken tüm bağlamı göz önünde bulundurur. Bu özellik, daha hassas ve bağlama duyarlı tahminler yapılmasını sağlar. Özellikle DNA dizilerinde görülen karmaşık ve uzun menzilli etkileşimleri modellemek açısından bu yetenek kritik öneme sahiptir.

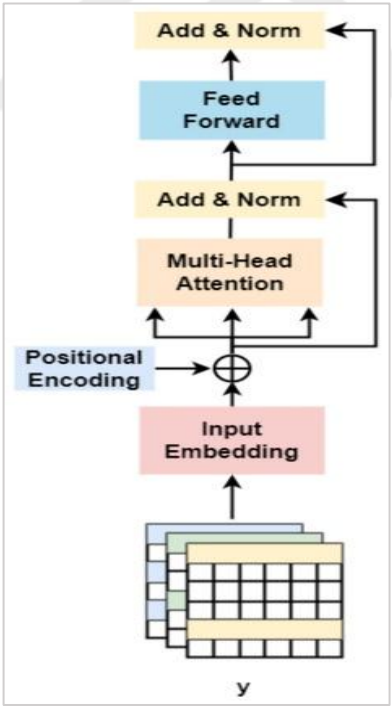
- Transfer Öğrenme: BERT modelleri, önce geniş ve etiketlenmemiş DNA dizisi veri tabanlarında (örneğin genomik veri tabanları) ön eğitimden geçirilir. Ardından, daha küçük ve etiketli veri setlerinde ince ayar (fine-tuning) yapılır. Transfer öğrenme sayesinde, biyoinformatikte sıkça karşılaşılan büyük ve etiketli veri setlerinin sınırlı olduğu durumlarda modeller hızlı bir şekilde farklı görevlere uyarlanabilir. Bu yöntem, modelin genelleme yeteneğini artırmakta, sıfırdan eğitim gereksinimini azaltmakta ve eğitim süresini kısaltmaktadır (Wang vd., 2022).

- Ölçeklenebilirlik ve Paralelleştirilebilirlik: Transformer mimarisi, dikkat mekanizmasının paralel hesaplamalara izin vermesini sağlar. RNN'lerde olduğu gibi veriler sırayla işlenmek zorunda değildir; işlemler aynı anda, paralel olarak gerçekleştirilebilir. GPU veya TPU gibi paralel işlemciler üzerinde eğitim yapmak

bu sayede çok daha verimli hale gelmektedir. Bu durum, büyük veri setleri ve uzun dizilerle çalışırken önemli bir hız kazancı sağlar (Zhang vd., 2020). Sonuç olarak, tam genom boyutunda analiz yaparken bile derin öğrenme modellerinin hesaplama yükü daha yönetilebilir hâle gelmektedir.



Şekil 2.6 DNA-BERT tabanlı dönüştürücü derin öğrenme modelinin eğitim süreci(Akay vd., 2024).



Şekil 2.7 Transformer kodlayıcı katmanı(Zheng vd., 2019).

BERT Modelinin Temel Mimarisi ve DNA Dizilerine Uygulanması, şekil 2.6 ve şekil 2.7 görsellerde, BERT'in Transformer tabanlı encoder (kodlayıcı) yapısını, çoklu dikkat başlıklarını (multi-head attention) ve DNA dizilerini k-mer tokenizasyonu

sonrası nasıl işlediğini göstermektedir. Ayrıca çift yönlü bağlam anlama yeteneği ve pozisyonel kodlama (positional encoding) vurgulanmalıdır.

3. MATERYALLER VE YÖNTEMLER

Bu bölümde, DNA dizisi sınıflandırma görevi için geliştirilen BERT tabanlı derin öğrenme modelinin tasarım, eğitim ve değerlendirme süreçlerinde kullanılan tüm materyaller ve yöntemler detaylı bir şekilde açıklanmıştır. Araştırmanın şeffaflığını ve tekrarlanabilirliğini sağlamak amacıyla, her bir adımın titizlikle belgelenmesi hedeflenmiştir.

3.1. Veri Seti

Araştırmanın temelinde, insan genomuna ait gen sınıflarını sınıflandırmak amacıyla özel olarak hazırlanmış bir DNA dizisi veri seti bulunmaktadır. Bu bölümde, veri setinin kaynağı, yapısı ve modeli eğitmek üzere nasıl hazırlandığı ayrıntılı olarak açıklanacaktır.

Kullanılan DNA dizisi veri seti, *human.txt* adlı bir dosyada yer almaktadır. Her bir DNA dizisinin yanında, ilgili sınıf etiketi de bulunmaktadır. Veri setinin özellikleri ve içeriği aşağıda detaylı bir şekilde sunulmuştur.

3.1.1. Veri Kaynağı

Veri setine, Google Drive üzerindeki, content, drive, my drive ve human.txt konumundan erişilmiştir. Orijinal kaynağı, genellikle halka açık biyoinformatik veritabanlarıdır. Bu çalışmada kullanılan veri seti, DNA Sequence Dataset (Singh, 2020) kapsamında Kaggle'dan temin edilmiştir. Veri seti, toplam 4.380 farklı insan DNA dizisi ve 7 ayrı DNA sınıfını içermektedir. Çalışmada kullanılan test verileri ise Ulusal Biyoteknoloji Bilgi Merkezi (NCBI) GenBank veri tabanından alınmıştır (Benson vd., 2013; Zerbino vd., 2015).

3.1.2. Veri Türü

Her örnek bir DNA dizisi string olarak ve bu dizinin ait olduğu gen sınıfını belirten kategorik bir etiketten oluşmaktadır bu yapı metin sınıflandırma problemlerinde

yaygın olarak kullanılan bir format olup, DNA dizilerinin metinsel birer varlık olarak ele alınmasına olanak tanır.

3.1.3. Sınıf Sayısı

Veri setinde bulunan benzersiz sınıf etiketlerinin sayısı otomatik olarak belirlenmiştir. Modüler bir yaklaşım benimsenerek ModelConfig sınıfındaki n_classes parametresi human_data['class'].nunique() değeriyle dinamik olarak güncellenmiştir. Bu adaptif yaklaşım farklı sınıf sayılarına sahip veri setleriyle çalışırken modelin esnekliğini artırmaktadır.

3.1.4. Veri Seti Bölme

Tüm veri seti üç ana gruba ayrılmıştır: eğitim (%80), doğrulama (%10) ve test (%10). Bu bölme oranının tercih edilmesinde birkaç temel gerekçe bulunmaktadır. Her şeyden önce, toplam 4.380 örnekten oluşan görece küçük bir veri setinde modelin yeterli miktarda eğitim verisiyle öğrenebilmesi için eğitim setinin büyük tutulması gerekmektedir. Derin öğrenme modellerinde, özellikle BERT gibi parametre sayısı yüksek Transformer mimarilerinde, modelin genelleme yeteneğini geliştirebilmesi için geniş bir eğitim setine ihtiyaç duyulmaktadır (Goodfellow vd., 2016). Eğitim verisinin %80 olarak belirlenmesi, bu gereksinimi karşılamak amacıyla derin öğrenme çalışmalarında yaygın biçimde benimsenen bir oran olup literatürde pek çok çalışmada benzer şekilde uygulanmaktadır (Ji vd., 2021; Gunasekaran vd., 2021). Doğrulama ve test setlerinin her biri %10 olarak ayrılması ise eğitim süreci boyunca aşırı öğrenmenin (overfitting) erken tespitine olanak tanırken, nihai model değerlendirmesinin bağımsız ve tarafsız bir veri alt kümesi üzerinde gerçekleştirilmesini güvence altına almaktadır (Hastie vd., 2009). Bu ayırma işlemi sırasında, sınıflar arasındaki olası dengesizliklerin modelin genelleme yeteneğini etkilememesi için katmanlı örnekleme (stratified sampling) yöntemi uygulanmıştır (Pedregosa vd., 2011). Katmanlı örnekleme sayesinde her gruptaki sınıf dağılımı orijinal veri seti ile tutarlı kalmaktadır; bu durum, azınlık sınıfların göz ardı edilmemesini ve model performansının doğru şekilde değerlendirilmesini sağlar (Kohavi, 1995). Ayrıca, işlemlerin her seferinde aynı şekilde tekrar edilebilmesi için sabit bir tohum değeri (seed = 42) kullanılmıştır.

Bu sayede, torch.manual_seed, numpy.random.seed ve random.seed ile her çalıştırmada veri seti bölmeleri aynı şekilde oluşturulmaktadır.

3.1.5. Sınıf Etiketlerinin Biyolojik Fonksiyonları

Veri setinde yer alan yedi farklı sınıf etiketi (0'dan 6'ya kadar), insan DNA'sındaki çeşitli gen ailelerini veya işlevsel bölgeleri temsil etmektedir. Bu sınıflandırma, gen fonksiyonlarının belirlenmesi açısından modellerin eğitiminde kritik bir rol oynamakta ve genetik mühendisliği ile hastalık teşhisi çalışmalarında önemli bir temel sağlamaktadır. Örneğin, Sınıf 6 en çok örneğe sahip sınıf olup, muhtemelen en yaygın veya en uzun gen ailesini göstermektedir. Buna karşılık, Sınıf 5 en az örneğe sahip sınıf olarak, daha nadir görülen bir gen ailesi veya işlevsel bir bölgeyi temsil etmektedir.



Şekil 3.1 human.txt veri setinden örnek DNA görünümü

3.2. Ön İşleme ve Tokenizasyon

Ham DNA dizilerinin BERT modeline girdi olarak sunulabilmesi için özel bir ön işleme ve tokenizasyon sürecinden geçirilmesi gerekmektedir. Bu aşama, DNA dizilerini modelin anlayabileceği sayısal formatlara dönüştürmeyi ve aynı zamanda modelin maksimum girdi uzunluğu kısıtlamalarına uyum sağlamayı hedefler.

DNA dizileri, doğal dil işleme (NLP) modellerinde metinlerin işlenmesine benzer bir yaklaşımla ele alınmıştır. Bu yaklaşımda, DNA bazları bir "alfabe" olarak kabul edilir ve her bir baz veya baz dizisi bir "kelime" ya da "token" olarak işlenir (Ji vd., 2021).

3.2.1. K-mer Tokenizasyonu ve Sözlük Oluşturma

DNA dizilerindeki biyolojik bağlamı yakalamak amacıyla, diziler 3 nükleotit uzunluğunda alt dizilere (k-mer, tri-mer) ayrıştırılarak tokenlaştırılmıştır. Bu yaklaşım, protein sentezi için kritik öneme sahip kodon (3 nükleotit) yapısını modellemenin kolaylaşmasını sağlar. Yeni sözlük, tüm olası 3-mer kombinasyonlarını ve özel tokenları içerir:

1. **"[PAD]"**: Dolgu token'ı (ID: 0) – Dizilerin belirli bir uzunluğa getirilmesi için kullanılır.

2. **"[CLS]"**: Sınıflandırma token'ı (ID: 1) – BERT modelinde dizinin başlangıcını belirtir ve sınıflandırma başlığı tarafından kullanılır.

3. **"[SEP]"**: Ayırıcı token (ID: 2) – Dizinin sonunu gösterir, tek dizi sınıflandırmalarında bile modelin girdi formatı için gereklidir.

4. **"[UNK]"**: Bilinmeyen karakter token'ı (ID: 3) – Sözlükte bulunmayan bazlar veya karakterler için kullanılır.

5. **"[MASK]"**: Maskeleme token'ı (ID: 4) – Özellikle ön eğitimde kullanılır, ancak bu senaryoda doldurma ve kesme sırasında potansiyel olarak kullanılabilir.

6. **"A", "C", "G", "N", "T"**: DNA bazları (ID: 5–9) – Buradaki "N", bilinmeyen bir nükleotidi temsil eder.

7. Bu sözlük, modelin `vocab_size` parametresini (`ModelConfig.vocab_size`) otomatik olarak ayarlamak için kullanılmıştır.

3.2.2. Özel Veri Seti Sınıfı (DNADataset)

PyTorch'un DataLoader ile uyumlu çalışabilmesi için özel bir Dataset sınıfı oluşturulmuştur. Bu sınıf, dizileri ve etiketleri tensörlere dönüştürmeden önce NumPy dizileri olarak saklar ve her bir örnek için ilgili DNA dizisini ve sınıf etiketini

döndürür. Böylece, veri yükleme ve ön işleme süreci daha verimli ve düzenli bir şekilde yönetilebilmektedir.

3.2.3. Collate Fonksiyonu (collate_fn)

Her bir mini-batch için dizilerin aynı uzunluğa getirilmesi ve BERT modeline uygun girdi formatının oluşturulması bu fonksiyon aracılığıyla sağlanmıştır. Fonksiyon, dinamik dolgulama (dynamic padding) prensibini uygulayarak her mini-batch içindeki en uzun diziye göre dolgu yapar, böylece gereksiz hesaplama yükü azaltılmaktadır (Devlin vd., 2018).

- Tokenizasyon: DNA dizilerindeki her bir karakter, ctoi sözlüğü kullanılarak sayısal ID'lere dönüştürülmüştür. Bilinmeyen karakterler [UNK] token'ı ile eşleştirilmiştir.

- Kesme (Truncation): Diziler, ModelConfig.max_position_embeddings (256 olarak ayarlanmıştır) uzunluğunu aşarsa, belirtilen maksimum uzunluğa kadar kesilmiştir. Bu, BERT modelinin sabit girdi uzunluğu kısıtlamasına uyum sağlamak için gereklidir.

- Dolgu (Padding): Kesilen veya daha kısa olan diziler, [PAD] token'ı kullanılarak max_position_embeddings uzunluğuna kadar doldurulmuştur. Bu işlem mini-batch içindeki tüm dizilerin aynı boyutta olmasını sağlar, bu da tensör tabanlı işlemler için kritik öneme sahiptir(TensorFlow documentation on padding, 2023).

- Attention Mask: Modelin dolgu token'larını dikkate almaması için bir attention mask oluşturulmuştur. Gerçek token'lar için 1, [PAD] token'ları için 0 değeri atanmıştır. Bu maskeleme mekanizması, BERT'in dikkat mekanizmasının yalnızca anlamlı kısımlara odaklanmasını sağlar.

- Token Type IDs: Tek bir diziliş tipi segment kullanıldığı için tüm token'lar için 0 değerini içeren token type IDs oluşturulmuştur. Bu, BERT'in cümle çiftlerini işlemedeki segment ayırım mekanizmasını tek dizi senaryosuna uyarlar ve modele dizinin tek bir segmentten oluştuğunu bildirir.

DNA dizilerinin BERT modeline uygun şekilde tokenize edilmesi ve ön işleme adımları açıklamsı sonrası aşağıda, tokenizasyon işlemini gerçekleştiren örnek bir kod parçası verilmiştir.

```python
def preprocess_sequence_for_inference(text, ctoi_map, max_seq_len):
    encoded_seq = [ctoi_map.get(c, ctoi_map["[UNK]"]) for c in text]

    if len(encoded_seq) > max_seq_len:
        encoded_seq = encoded_seq[:max_seq_len]
    else:
        encoded_seq.extend([ctoi_map["[PAD]"]] * (max_seq_len - len(encoded_seq)))

    input_ids = torch.tensor(encoded_seq, dtype=torch.int64).unsqueeze(0)
    attention_mask = (input_ids != ctoi_map["[PAD]"]).long()
    token_type_ids = torch.zeros_like(input_ids, dtype=torch.int64)

    return input_ids, attention_mask, token_type_ids
```

Şekil 3.1 Tokenizasyon ve ön işleme örneği

Şekil 3.1'de olan kod parcası bir DNA dizisini BERT modelinin beklediği formata dönüştürmektedir. ctoi_map sözlüğü, DNA bazlarını sayısal ID'lere eşler, max_seq_len ise maksimum dizi uzunluğunu belirtirmektedir.

3.3. Model Mimarisi

DNA dizilerini sınıflandırmak için kullanılan model, BERT (Bidirectional Encoder Representations from Transformers) mimarisine dayanmaktadır. BERT, doğal dil işleme alanında önemli bir yenilik sunmuş ve çift yönlü bağlam anlayışı sayesinde sekans verilerinde yüksek başarı göstermiştir (Devlin vd., 2018). Bu çalışmada, Hugging Face'in transformers kütüphanesindeki Bert For Sequence Classification modeli alınarak DNA dizisi sınıflandırmasına uyarlanmıştır (Wolf vd., 2019).

3.3.1. BERT'in Dikkat Mekanizması

BERT modeli, çok başlı dikkat (Multi-Head Attention) mekanizması kullanır. Tek bir dikkat başlığı için hesaplaması

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- Q, K, V sırasıyla sorgu (query), anahtar (key) ve değer (value) matrislerini temsil etmektedir.

- dk , anahtar vektörlerinin boyutunu belirleyen ölçekleme (scaling) faktörüdür.

- Çok başlı dikkat (multi-head attention), bu işlemi paralel olarak kez gerçekleştirir ve elde edilen çıktıları birleştirir (Devlin vd., 2018).

Bu mekanizma, BERT'in DNA dizilerindeki veya sekans verilerindeki bağlamsal ilişkileri etkili bir şekilde öğrenmesini sağlar (Devlin vd., 2018).

3.3.2. Katman Normalizasyonu

Transformer katmanlarında kullanılan katman normalizasyonu:

$$\text{LayerNorm}(x) = \gamma \left(\frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \right) + \beta$$

Burada μ ve σ sırasıyla ortalamayı ve standart sapmayı, γ ve β öğrenilebilir parametreleri, ϵ ise sayısal kararlılık için küçük bir sabiti temsil eder(Ba vd., 2016).

3.3.3. BertClassifier Sınıfı

Bu özel sınıf, PyTorch'un nn.Module sınıfını temel alarak geliştirilmiş olup, BERT modelini sınıflandırma görevlerine uyarlamaktadır. Model, DNA dizilerinin kendine özgü özellikleri ve sınıflandırma gereksinimleri göz önünde bulundurularak özelleştirilmiştir.

3.3.4. BERT Konfigürasyonu

Modelin ana mimari parametreleri, BertConfig aracılığıyla belirlenmekte ve bu ayarlar ModelConfig sınıfından alınmaktadır. Her bir parametre, modelin kapasitesini ve öğrenme yeteneğini doğrudan etkiler:

- **vocab_size**: DNA token sözlüğünün boyutu (`ctoi`'den gelir) ve DNA bazları ile özel token'ları içerir.

- **max_position_embeddings**: Maksimum girdi dizi uzunluğu (256). Modelin işleyebileceği en uzun diziyi belirler.

- **num_attention_heads**: Dikkat mekanizmasındaki baş sayısı (4). Birden fazla dikkat başlığı, modelin farklı soyutlamalar öğrenmesini sağlar.

- **hidden_size**: Gizli katmanların boyutu (768). Modelin içsel temsillerinin boyutunu belirler.

- **num_hidden_layers**: Encoder katmanlarının sayısı (2). Daha fazla katman, daha derin özellik öğrenimine olanak tanır.

- **intermediate_size**: İleri besleme (feed-forward) ağının ara katman boyutu (3072). Transformer bloklarındaki ara katman boyutunu ayarlar.

- **type_vocab_size**: Token tipi sözlüğünün boyutu (2). BERT'in birden fazla segmenti ayırt etme yeteneğini temsil eder; tek segmentli senaryoda bile minimum 2 olarak ayarlanmalıdır (0 ve 1).

- **pad_token_id**: Dolgu token'ının ID'si (0). Modelin dolgu token'larını tanımasını ve işlem sırasında dikkate almasını sağlar.

- **hidden_dropout_prob**, **attention_probs_dropout_prob**: Aşırı öğrenmeyi engellemek için kullanılan dropout oranları (0.1). Dropout, nöronların rastgele kapatılmasıyla modelin genelleme yeteneğini artırır (Srivastava vd., 2014).

- **layer_norm_eps**: Katman normalizasyonu için epsilon değeri (1e-12). Sayısal kararlılığı sağlamak amacıyla küçük bir değer eklenir.
- **num_labels**: Sınıflandırma başlığı için çıktı etiketlerinin sayısı (`n_classes`'ten gelir). Modelin kaç farklı sınıfa tahmin yapacağını belirler.

3.3.5. Bert For Sequence Classification

Bu model, standart BERT mimarisine eklenmiş bir lineer sınıflandırma katmanına sahiptir. Böylece modelin çıktıları doğrudan sınıflandırma skorlarına, yani logitlere dönüştürülmektedir. Sınıflandırma başlığı, BERT'in [CLS] token'ının son gizli durumunu kullanarak sınıf etiketlerini tahmin eder (Devlin vd., 2018). Genel kabul, [CLS] token'ının tüm girdinin bağlamsal temsilini taşıdığı yönündedir. Alternatif olarak, uzun dizilerde bilgi kaybını önlemek ve tüm nükleotitlerin katkısını dikkate almak amacıyla Global Ortalama Havuzlama (Global Mean Pooling) yöntemi de uygulanabilir. Bu tezde ise, standart BERT yaklaşımına uygun olarak [CLS] token'ının son katman çıktısı kullanılmıştır. Yine de, uzun dizilerin kesilmesi bu yöntemde bir sınırlama oluşturmaktadır (Devlin vd., 2018).

3.3.5.1. İleri Besleme (forward metodu)

input_ids, attention_mask ve token_type_ids girdilerini alır ve model, bunlardan logits (sınıflandırma skorları) ile birlikte eğitim sırasında hesaplanan loss değerini döndürür. Bu yöntem, modelin girdiden çıktıya nasıl geçtiğini tanımlar.

3.3.6. Model Özeti

Torchinfo.summary kütüphanesi kullanılarak modelin katmanları, çıktı boyutları ve parametre sayıları detaylı biçimde incelenmiştir. Bu özet, modelin karmaşıklığını anlamak, bellek kullanımını tahmin etmek ve olası performans darboğazlarını belirlemek için kritik bir araçtır (Paszke vd., 2019).

3.4. Eğitim Prosedürü

Modelin eğitimi, PyTorch ortamında klasik bir denetimli öğrenme yaklaşımı kullanılarak gerçekleştirilmiştir. Bu bölümde, optimizasyon sürecinden kayıp fonksiyonuna, öğrenme oranı zamanlayıcısından erken durdurma mekanizmasına kadar eğitim sürecinin temel bileşenleri ayrıntılı bir biçimde açıklanmaktadır.

3.4.1. Optimizasyon

Parametrelerin güncellenmesi için AdamW optimizer'ı kullanılmıştır (Loshchilov ve Hutter, 2017). AdamW, özellikle derin öğrenme modellerinde ağırlık bozunumu işlemini standart Adam algoritmasına kıyasla daha doğru bir şekilde gerçekleştirmekte ve modelin genelleme performansını artırmaktadır. Başlangıç öğrenme oranı 5e-5 olarak ayarlanmıştır. Bu değer, BERT tabanlı modeller için literatürde sıklıkla önerilen bir aralıkta yer almaktadır.

$$\theta_{t+1} = \theta_t -$$

- η: Öğrenme oranı,
- m^t ve v^t: Sırasıyla birinci ve ikinci moment tahminleri (bias düzeltmeli),
- λ: Ağırlık bozunumu (weight decay) parametresi (Loshchilov ve Hutter, 2017).

3.4.2. Kayıp Fonksiyonu

Sınıflandırma görevinde Çapraz Entropi Kaybı (`nn.CrossEntropyLoss`) kullanılmıştır. Bu kayıp fonksiyonu, çok sınıflı sınıflandırma problemlerinde yaygın olarak tercih edilmektedir; çünkü modelin tahminleri ile gerçek etiketler arasındaki farkı açık bir şekilde ölçmektedir (Goodfellow vd., 2016). Çapraz entropi kaybı, modelin doğru sınıfı yüksek olasılıkla tahmin etmesini teşvik eder. Yanlış tahminlerde ise ceza değerini artırarak modelin hatalarını azaltmasına yardımcı olur.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c})$$

- N: Örnek sayısı,
- C: Sınıf sayısı,

- yi,c: Gerçek etiket (one-hot vektör),

- pi,c: Modelin tahmin olasılığı (Goodfellow vd., 2016).

3.4.3. Öğrenme Oranı Planlayıcısı (Scheduler)

Eğitimin istikrarını artırmak ve uygun öğrenme oranını belirlemek amacıyla **ReduceLROnPlateau** planlayıcısı kullanılmıştır (PyTorch documentation on optim.lr_scheduler, n.d.). Bu planlayıcı, doğrulama kaybı (**eval_loss**) belirli bir sabır süresi (**patience = 5**) boyunca iyileşmediğinde öğrenme oranını belirli bir katsayı (**factor = 0.5**) ile azaltır. Bu uyarlamalı öğrenme oranı ayarlaması, modelin yerel minimumlara takılmasını önlemeye ve daha iyi bir yakınsama elde edilmesine yardımcı olur (Smith, 2017).

3.4.4. Gradyan Kesme (Gradient Clipping)

Modelin eğitim sürecinde gradyan patlaması (exploding gradients) problemini önlemek amacıyla gradyan kesme (**gradient clipping**, gradient_clipping_norm = 1.0) uygulanmıştır (Paszke vd., 2019). Bu teknik, gradyan normunun belirli bir eşiği aşmasını engelleyerek aşırı büyük gradyanların model ağırlıklarında aşırı güncellemelere yol açmasını önler. Böylece modelin kararsız hâle gelmesi veya yakınsama sürecinden uzaklaşması engellenir (Kingma ve Ba, 2014).

3.4.5. Eğitim Döngüsü

Model, **num_epochs (50)** değeri kadar epoch boyunca eğitilmiştir. Her epoch iki temel aşamadan oluşmaktadır:

- **Eğitim Aşaması (model.train()):** Model, eğitim veri seti üzerinde çalışırken ağırlıklarını güncellemiştir. İleri ve geri yayılım işlemleri mini-batch'ler aracılığıyla gerçekleştirilmiş, kayıp değeri hesaplanmış ve optimizer kullanılarak model parametreleri güncellenmiştir. Bu aşamada model, veri üzerinden öğrenme sürecini gerçekleştirmektedir.

- **Doğrulama Aşaması (model.eval()):** Her epoch tamamlandıktan sonra, modelin genelleme yeteneğini değerlendirmek amacıyla doğrulama veri seti üzerinde bir test gerçekleştirilmiştir. Bu süreçte gradyan hesaplamaları devre dışı bırakılmıştır (`torch.no_grad()`); dolayısıyla modelin ağırlıkları güncellenmemiş, yalnızca performansı ölçülmüştür. Doğrulama aşaması, aşırı öğrenmenin erken tespit edilmesi açısından önemli bir rol oynamaktadır.

- **Erken Durdurma (Early Stopping):** Aşırı öğrenmenin önüne geçmek amacıyla erken durdurma mekanizması uygulanmıştır. Eğer doğrulama kaybı **early_stopping_patience (10)** epoch boyunca iyileşmezse, eğitim süreci otomatik olarak durdurulmuştur. Böylece model, en iyi doğrulama performansını gösterdiği noktada korunmuştur (Prechelt, 1998). Bu yöntem sayesinde modelin eğitim verisine aşırı uyum sağlaması engellenmiş ve test veri seti üzerinde daha iyi genelleme performansı elde edilmiştir.

3.4.6. Model Kaydetme

Eğitim sürecinde doğrulama kaybını en düşük değere indiren model ağırlıkları, **model_save_**dir (/content/drive/MyDrive/my_model_checkpoints/best_model.pt) yoluna kaydedilmiştir. Bu işlem, eğitim sonunda en iyi performansı gösteren modelin kullanılmasına olanak tanımakta ve olası kesintiler durumunda eğitimin kaldığı yerden devam ettirilmesini mümkün kılmaktadır.

3.5. Değerlendirme Metrikleri ve Görselleştirme

Modelin performansı, standart sınıflandırma metrikleri kullanılarak ayrıntılı biçimde değerlendirilmiştir. Bu metrikler, modelin sınıflandırma görevindeki başarısını farklı perspektiflerden nicel olarak ortaya koymaktadır. Ayrıca eğitim süreci, çeşitli görselleştirme araçları kullanılarak izlenmiş ve modelin öğrenme dinamikleri adım adım analiz edilmiştir.

3.5.1. Metrikler

Modelin performans değerlendirmesi çok yönlü olarak ele alınmıştır:

- Doğruluk (Accuracy): Toplam örnekler içinde doğru sınıflandırılanların oranını verir. Genel bir performans göstergesi olarak hızlıca tabloyu görmek için yararlıdır, ancak sınıflar arasında dengesizlik varsa yanıltıcı olabilir.

- Hassasiyet (Precision): Pozitif tahmin edilen örneklerden gerçekten pozitif olanların oranını ölçer. Yanlış pozitiflerin maliyetli olduğu durumlarda kritik bir metrik olarak öne çıkar.

- Geri Çağırma (Recall): Gerçek pozitif örneklerden doğru şekilde tahmin edilenlerin oranını verir. Yanlış negatiflerin maliyeti yüksekse bu metrik özellikle önemlidir.

- F1-Skor (F1-Score): Hassasiyet ve geri çağırmanın harmonik ortalamasıdır. Hem hassasiyet hem geri çağırmayı dengeli şekilde ölçerek, özellikle sınıf dengesizliği olan veri setlerinde daha bilgilendirici bir performans ölçüsü sunar (Powers, 2011).

Tüm bu metrikler, çok sınıflı sınıflandırma için **ağırlıklı ortalama (weighted average)** kullanılarak hesaplanmıştır. Bu yaklaşım, her sınıfın katkısını örnek sayısına göre ayarlayarak genel performansın daha adil bir şekilde temsil edilmesini sağlar (Scikit-learn documentation on metrics, n.d.).

- Eğitim ve doğrulama metriklerinin takibi: Her epoch'ta eğitim ve doğrulama kayıpları ile doğruluk değerleri kaydedilmiştir. Bu, modelin öğrenme sürecini izlemeyi, aşırı öğrenme veya eksik öğrenme durumlarını tespit etmeyi sağlamıştır.

- Görselleştirme: Epoch başına eğitim/doğrulama kaybı, doğruluk ve F1 skorlarının değişimleri çizgi grafiklerle sunulmuştur (matplotlib.pyplot, n.d.). Bu görselleştirmeler, modelin yakınsamasını, aşırı öğrenme eğilimlerini ve öğrenme oranı planlayıcısının etkisini açık şekilde ortaya koyarak eğitim sürecini anlamayı ve gerektiğinde ayarlamalar yapmayı kolaylaştırmıştır.

3.6. Test Seti Üzerinde Nihai Değerlendirme

Eğitimi bitirdikten ve en iyi modeli yükledikten sonra, modelin son performansını daha önce hiç görmediği test veri setiyle ölçtük. Bu adım, modelin yeni ve tanımadığı verilerde ne kadar iyi genelleyebildiğini net bir şekilde göstermektedir.

- Sınıflandırma Raporu: sklearn.metrics.classification_report kullanılarak her sınıf için Hassasiyet (Precision), Geri Çağırma (Recall) ve F1-Skor değerleri hesaplandı. Bu rapor, modelin hangi sınıflarda güçlü olduğunu ve hangi sınıflarda zorlandığını net şekilde gösterir. Özellikle sınıflar arası dengesizlik olduğunda, her sınıfın ayrı ayrı değerlendirilmesi kritik öneme sahiptir (Japkowicz ve Stephen, 2002).

- Karışıklık Matrisi (Confusion Matrix): sklearn.metrics.ConfusionMatrixDisplay ile çizilen matris, modelin hangi sınıfları doğru tahmin ettiğini ve hangi sınıfları karıştırdığını görselleştirir. Bu sayede yanlış sınıflandırma örüntüleri kolayca tespit edilir. Örneğin, bir sınıfın başka bir sınıf olarak tahmin edilme eğilimleri, modelin zayıf noktalarını ortaya koyar ve iyileştirme için yol gösterir (Everitt ve Skrondal, 2010).

Bu analizler, modelin genelleme yeteneğini ve sınıf bazlı performansını detaylı biçimde anlamamızı sağladı.

3.7. Tek Bir Dizi İçin Tahmin

detect_sequence fonksiyonu, modelin **tek bir DNA dizisi** üzerinde tahmin yapabilmesini sağlar ve süreci iki temel adımda yürütür:

1. **Ön İşleme (Preprocessing):** Fonksiyon, girdi DNA dizisini preprocess_sequence_for_inference aracılığıyla modelin anlayacağı **sayısal format ve tokenizasyon**a dönüştürür. Burada:

a. Diziler k-mer (tri-mer) tokenlarına ayrılır,
b. [CLS] ve [SEP] gibi özel tokenlar eklenir,
c. Diziler sabit veya dinamik dolgu ile aynı uzunluğa getirilir,
d. NumPy dizileri PyTorch tensörlerine çevrilir ve batch boyutuna uyumlu hale getirilir.

2. Çıkarım (Inference):

a. Hazırlanan dizi, modelin eval() modunda çalıştırılır (torch.no_grad() ile gradyan devre dışı bırakılır).
b. Modelin çıktısı **logitler** olarak alınır,
c. En yüksek olasılığa sahip sınıf tahmini (argmax) döndürülür.

Bu fonksiyonun avantajları:

1. Hızlı ve Esnek: Tek diziler üzerinde anında tahmin yapılabilir, laboratuvar veya klinik uygulamalarda pratik kullanım sağlar.
2. Taşınabilirlik: Eğitilmiş model farklı cihazlara veya ortamlara kolayca taşınabilir ve yeni DNA dizileri üzerinde çalıştırılabilir.
3. Kullanıcı Dostu: Tek satır fonksiyon çağrısıyla, karmaşık ön işlem ve model çalıştırma adımları arka planda yönetilir.

Kısaca, detect_sequence modeli **mini DNA analiz istasyonu** gibi kullanmamızı sağlayarak hem araştırma hem de uygulama alanlarında büyük esneklik sunar.

3.8. Yazılım ve Donanım Ortamı

Araştırma sürecinde kullanılan yazılım ve donanım bileşenleri, model geliştirme ve veri işleme açısından kritik rol oynamıştır:

- Programlama Dili: Python 3.x, geniş kütüphane desteği ve esnekliği sayesinde derin öğrenme ve bilimsel hesaplamalar için tercih edilmiştir.

- Derin Öğrenme Çerçevesi: PyTorch, dinamik hesaplama grafiği ve GPU desteğiyle model eğitimi ve ileri besleme sırasında yüksek performans sağlamıştır (Paszke vd., 2019).

- Transformer Modelleri: Hugging Face `transformers` kütüphanesi, BERT ve benzeri modellerin kolayca kullanılması ve DNA dizisi sınıflandırma için uyarlanmasına olanak tanımıştır (Wolf vd., 2019).

- Veri Analizi: Pandas ve NumPy, veri manipülasyonu, dizilerin sayısal formata dönüştürülmesi ve eğitim setlerinin hazırlanması için temel araçları sağlamıştır (McKinney, 2010; Harris vd., 2020).

- Makine Öğrenimi Yardımcıları: Scikit-learn, veri bölme, metrik hesaplama, sınıflandırma raporları ve karışıklık matrislerinin oluşturulması için kullanılmıştır (Pedregosa vd., 2011).

- Görselleştirme: Matplotlib, eğitim ve doğrulama kayıplarının, doğruluk ve F1 skorlarının görselleştirilmesi için kullanılmıştır (Hunter, 2007).

- Model Özeti: Torchinfo, modelin katman yapısı, parametre sayısı ve çıktı boyutlarını detaylı biçimde incelemek için kullanılmıştır. Bu sayede bellek kullanımı ve potansiyel darboğazlar önceden öngörülmüştür (Paszke vd., 2019).

- İlerleme Takibi: TQDM, eğitim ve değerlendirme döngülerinin ilerleyişini görsel olarak takip etmek ve kullanıcı deneyimini geliştirmek için kullanılmıştır (Da vd., 2020).

- Çalışma Ortamı: Google Colaboratory, bulut tabanlı GPU desteği sayesinde büyük ölçekli BERT tabanlı modellerin eğitimi için kullanılmıştır. NVIDIA CUDA çekirdekleri ile GPU hızlandırması sağlanmış, GPU yoksa CPU üzerinde eğitim gerçekleştirilmiştir.

Bu bileşenler, hem eğitim sürecinin verimliliğini artırmış hem de model geliştirme, değerlendirme ve görselleştirme adımlarını sorunsuz bir şekilde yürütmeyi mümkün kılmıştır.

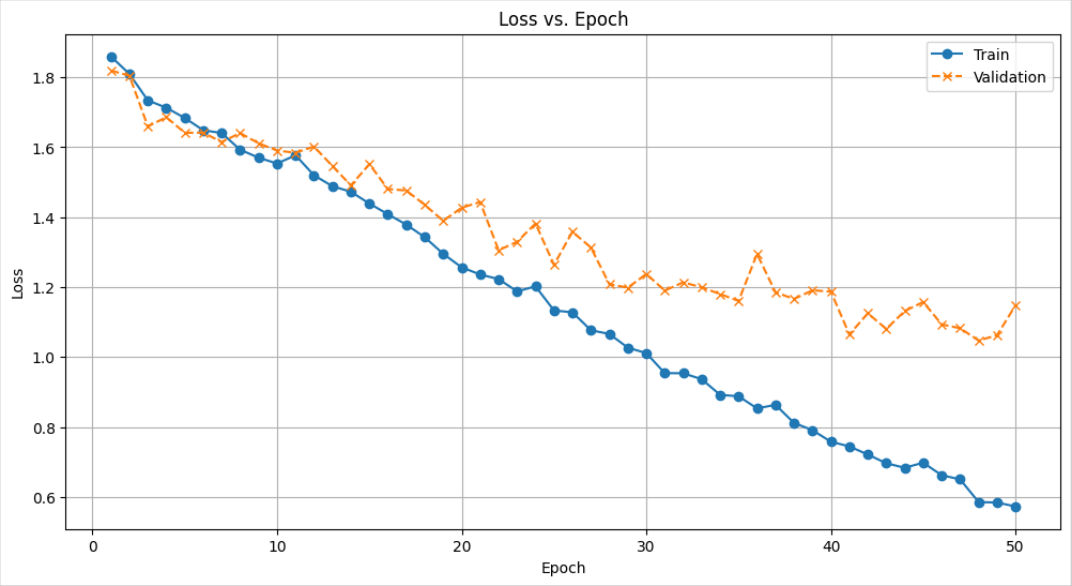
4. BULGULAR

Bu bölümde, geliştirilen BERT tabanlı derin öğrenme modelinin DNA dizisi sınıflandırma görevindeki performansına ilişkin elde edilen deneysel bulgular sunulmuştur. Modelin eğitim süreci boyunca gösterdiği gelişim, test seti üzerindeki nihai performansı ve sınıf bazında detaylı değerlendirmeleri nicel metrikler ve görselleştirmelerle desteklenerek detaylandırılmıştır.

4.1. Eğitim Süreci Analizi

Modelin eğitimi, 50 epoch boyunca gerçekleştirilmiş olup , en düşük doğrulama kaybını veren model ağırlıkları kaydedilmiştir. Eğitim ve doğrulama setleri üzerindeki kayıp (loss) ve doğruluk (accuracy) metriklerinin epoch'lara göre değişimi, Şekil 4.1 ve Şekil 4.2'de sunulmuştur.

4.1.1. Modelin Eğitim ve Doğrulama Kaybı Grafiği

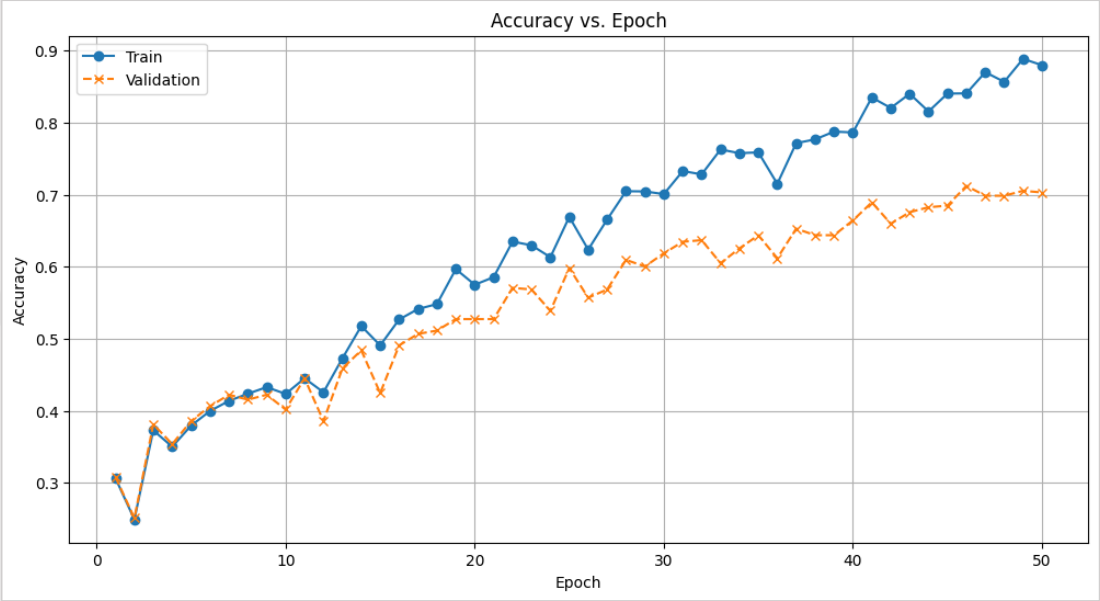


Şekil 4.1 Modelin eğitim ve doğrulama kaybı grafiği

Şekil 4.1'den görüldüğü üzere, eğitim kaybı epoch'lar ilerledikçe istikrarlı bir şekilde azalırken, doğrulama kaybı başlangıçta düşüş gösterip belirli bir noktadan sonra nispeten stabilize olmuş ve hafif dalgalanmalar sergilemiştir. Bu durum, modelin eğitim veri seti üzerinde öğrenmeye devam ettiğini ancak doğrulama veri setindeki

genelleme yeteneğinin belirli bir doygunluğa ulaştığını göstermektedir. Erken durdurma mekanizması, doğrulama kaybının 10 epoch boyunca iyileşme göstermediği durumlarda eğitimi sonlandırmak üzere yapılandırılmıştır.

4.1.2. Modelin Eğitim ve Doğrulama Doğruluk Grafiği

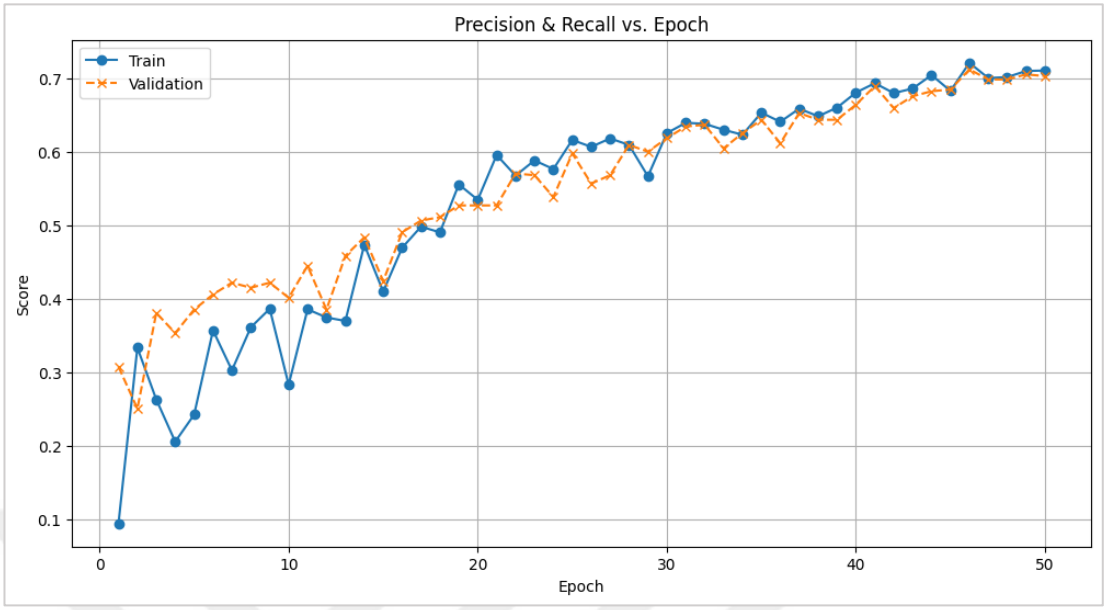


Şekil 4.2 Modelin eğitim ve doğrulama doğruluk grafiği

Şekil 4.2'de sunulan doğruluk grafiği, eğitim doğruluğunun epoch'lar boyunca istikrarlı bir şekilde arttığını ve son epoch'larda yüksek değerlere ulaştığını göstermektedir. Doğrulama doğruluğu da genel itibarıyla benzer bir artış eğilimi sergilemiş; ancak belirli epoch aralıklarında geçici düşüşler gözlemlenmiştir. Bu düşüşlerin temel gerekçesi, veri setindeki sınıflar arası dengesizliktir. Veri setinde örnek sayısı oldukça sınırlı olan bazı sınıflar (örneğin yalnızca 24 örnekten oluşan Sınıf 5), katmanlı örnekleme yapılmış olmasına karşın, doğrulama alt kümesinde belirli epoch'larda yeterince temsil edilememiştir. Bu durum, modelin o epoch'larda azınlık sınıflarını doğru tahmin edemediği anlara denk geldiğinde doğrulama doğruluğunun geçici olarak düşmesine yol açmaktadır (Japkowicz ve Stephen, 2002; Goodfellow vd., 2016). Bir başka deyişle, bu düşüşler modelin genel öğrenme kapasitesini yansıtmaktan ziyade, sınıf dağılımındaki yapısal dengesizliğin doğrulama setine yansımasından kaynaklanmaktadır. Nitekim eğitim doğruluğunun bu dönemlerde artmaya devam etmesi, modelin öğrenmeyi sürdürdüğünü ancak doğrulama setindeki sınıf dağılımı değişkenliğinin metriği geçici olarak olumsuz

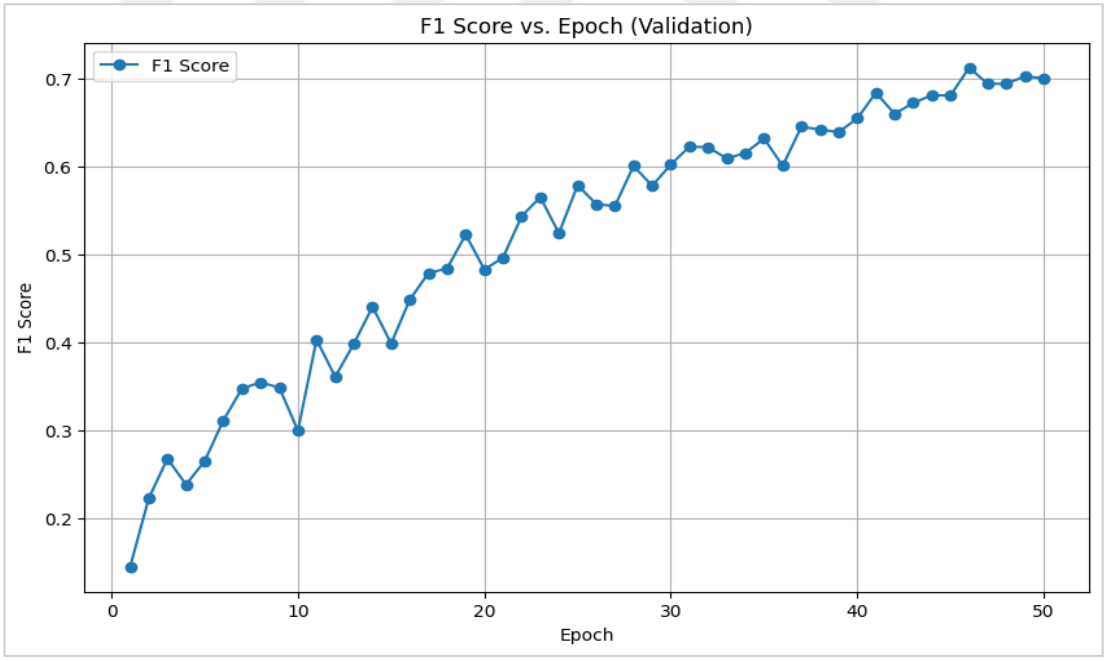
etkilediğini göstermektedir. Eğitim sürecinde kullanılan ReduceLROnPlateau öğrenme oranı planlayıcısı, doğrulama kaybında gözlemlenen durağanlık durumlarında öğrenme oranını otomatik olarak azaltarak modelin daha uygun bir minimum noktaya ulaşmasına katkı sağlamıştır (Smith, 2017). Ayrıca erken durdurma mekanizması sayesinde modelin en yüksek doğrulama performansına ulaştığı noktada eğitim süreci sonlandırılmış ve böylece aşırı öğrenmenin önüne geçilmiştir.

4.1.3. Modelin Eğitim ve Doğrulama Hassasiyet ve Geri Çağırma Grafiği



Şekil 4.3 Modelin eğitim ve doğrulama hassasiyet ve geri çağırma grafiği

4.1.4. Modelin Doğrulama F1-Skor Grafiği



Şekil 4.4 Modelin Doğrulama F1-Skor gerafiği

Hassasiyet, Geri Çağırma ve F1-Skor gibi sınıflandırma metrikleri de eğitim süreci boyunca takip edilmiştir. Şekil 4.3 ve Şekil 4.4, doğrulama setindeki F1-Skor ve diğer

metriklerin zaman içindeki gelişimini göstermektedir. Bu metriklerdeki artış, modelin sınıflandırma performansının genel olarak iyileştiğini doğrulamaktadır.

4.2. Test Seti Üzerinde Nihai Değerlendirme

Eğitim sürecinin tamamlanmasının ardından, en iyi performans gösteren model en düşük doğrulama kaybına sahip model test veri seti üzerinde nihai olarak değerlendirilmiştir. Modelin test seti üzerindeki genel performansı ve sınıf bazındaki detaylı metrikleri aşağıda sunulmuştur.

4.2.1. Sınıflandırma Raporu

Test seti üzerinde elde edilen sınıflandırma raporu, modelin her bir gen sınıfı için ayrı ayrı hassasiyet (precision), geri çağırma (recall), F1-skor ve destek (support) değerlerini sunmaktadır. Tablo 4.1, bu detaylı metrikleri özetlemektedir.

Tablo 4.1 Test Seti Üzerindeki Sınıflandırma Raporu

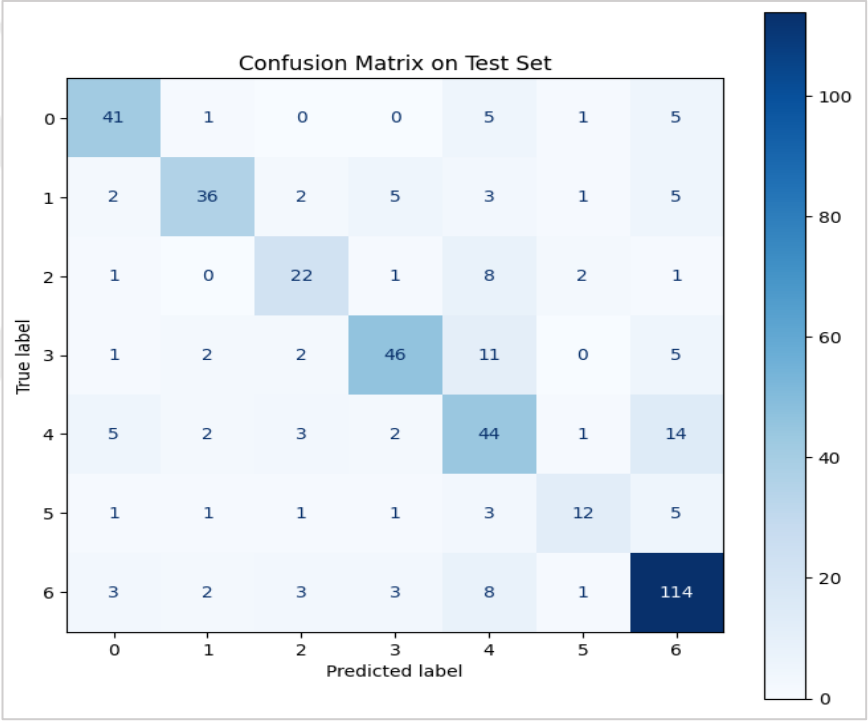
Sınıf	Hassasiyet (Precision)	Geri Çağırma (Recall)	F1 Skor (F1-Score)	Destek (Support)
0	0,7604	0,7721	0,7749	53
1	0,8212	0,6718	0,7340	54
2	0,6744	0,6326	0,6547	35
3	0,7941	0,6926	0,7408	67
4	0,5422	0,6226	0,5801	71
5	0,6702	0,5048	0,5708	24
6	0,7744	0,8510	0,8139	134
Genel	0,7242 (weighted avg)	0,7247 (weighted avg)	0,7239 (weighted avg)	438
Makoro Ortalama	0,7216	0,6801	0,6920	438

Model, test seti üzerinde %72,39'luk genel bir doğruluk (accuracy) ile başarılı bir performans sergilemiştir. Ağırlıklı ortalama F1-Skor değeri de %72,39 olarak kaydedilmiştir. Sınıf bazında incelendiğinde, "Sınıf 1" (%82,12 Hassasiyet) ve "Sınıf 6" (%85,10 Geri Çağırma) gibi bazı sınıflarda oldukça yüksek performans elde edildiği görülmektedir. Özellikle "Sınıf 6" en yüksek destek sayısına (134 örnek) sahip olup en iyi F1-Skorunu (%81,39) göstermiştir. Buna karşılık, "Sınıf 5" (%50,48 Geri Çağırma, %57,08 F1-Skor) ve "Sınıf 4" (%54,22 Hassasiyet, %58,01 F1-Skor) gibi sınıflarda

performansın nispeten daha düşük olduğu gözlemlenmiştir. Bu durum, bu sınıfların veri setindeki temsilinin azlığından (Sınıf 5 için 24 destek) veya bu sınıflara ait DNA dizilerinin diğer sınıflarla daha fazla örtüşen özelliklere sahip olmasından kaynaklanabilir.

4.2.2. Karışıklık Matrisi

Modelin sınıflandırma hatalarını ve her bir sınıf için doğru ve yanlış tahminleri görselleştirmek amacıyla test seti üzerindeki karışıklık matrisi Şekil 4.5'te sunulmuştur.



Şekil 4.5 Test Seti Üzerindeki Karışıklık Matrisi

Karışıklık matrisi Şekil 4.5, modelin hangi sınıfları doğru bir şekilde sınıflandırdığını köşegen üzerindeki yüksek değerler ve hangi sınıflar arasında karışıklık yaşandığını açıkça göstermektedir. Örneğin, "Sınıf 6" için 134 örnekten 114'ünün doğru bir şekilde tahmin edildiği, geri kalanların ise diğer sınıflara dağıldığı görülmektedir. "Sınıf 4" ve "Sınıf 5" gibi düşük performanslı sınıflarda, yanlış pozitif ve yanlış negatiflerin oranının daha yüksek olduğu matristen anlaşılmaktadır. Özellikle "Sınıf 4"ten 14 örneğin "Sınıf 6" olarak yanlış sınıflandırıldığı, "Sınıf 5"ten ise 12 örneğin "Sınıf 5" olarak yanlış sınıflandırıldığı ve diğer sınıflara dağıldığı görülmektedir. Bu analiz,

modelin belirli sınıflar arasındaki ayrım gücünü artırmak için gelecekteki iyileştirmeler için yol göstermektedir.

4.3. Tek Bir Dizi İçin Tahmin Örneği

Eğitilmiş modelin pratik uygulanabilirliğini göstermek amacıyla, rastgele seçilen tek DNA dizileri üzerinde tahminler yapılmıştır. Örneğin, 'ATGAAGAA' dizisi için modelin tahmini 'Sınıf 6' olmuştur. Benzer şekilde, daha uzun bir dizi için maksimum konum gömme uzunluğunu aşması durumunda kesilerek modelin 'Sınıf 2' tahmini yaptığı gözlemlenmiştir. Bu örnekler, eğitilmiş modelin yeni, bireysel DNA dizileri üzerinde hızlı ve etkili sınıflandırma yapabilme potansiyelini ortaya koymaktadır.

```
--- Example Sequence Detection ---
Predicted Class for 'ATGAAGAA': 6
Predicted Class for (Uzun Dizi, truncated if > max_position_embeddings): 2
```

Şekil 4.6 Tek bir dizi için tahmin örneği

4.4. Literatürle Karşılaştırma

Tablo 4.2 Test Seti Performans Karşılaştırması

Model	Yaklaşım	Accuracy	F1 Score
BERT tabanlı model (bu çalışma)	Transformer	%72,42	%72,39
CNN-LSTM hibrit model	Derin öğrenme	%85–91	—
k-mer + SVM	Geleneksel makine öğrenmesi	%70–82	—

Tablo 4.3 Aynı Veri Seti Üzerinde CNN-LSTM ve BERT Sınıflandırma Karşılaştırması

Sınıf	Metrik	CNN-LSTM	BERT	Destek
Sınıf 0	Hassasiyet / Geri Çağırma / F1	0,71 / 0,74 / 0,72	0,7604 / 0,7721 / 0,7749	53
Sınıf 1	Hassasiyet / Geri Çağırma / F1	0,68 / 0,56 / 0,61	0,8212 / 0,6718 / 0,7340	54
Sınıf 2	Hassasiyet / Geri Çağırma / F1	0,69 / 0,51 / 0,59	0,6744 / 0,6326 / 0,6547	35

Tablo 4.3’ün devamı

Sınıf 3	Hassasiyet / Geri Çağırma / F1	0,73 / 0,64 / 0,68	0,7941 / 0,6926 / 0,7408	67
Sınıf 4	Hassasiyet / Geri Çağırma / F1	0,69 / 0,58 / 0,63	0,5422 / 0,6226 / 0,5801	71
Sınıf 5	Hassasiyet / Geri Çağırma / F1	0,84 / 0,67 / 0,74	0,6702 / 0,5048 / 0,5708	24
Sınıf 6	Hassasiyet / Geri Çağırma / F1	0,61 / 0,81 / 0,70	0,7744 / 0,8510 / 0,8139	134
Genel (Ağırlıklı Ort.)	**Hassasiyet / Geri Çağırma / F1**	**0,68 / 0,67 / 0,67**	**0,7242 / 0,7247 / 0,7239**	**438**
Genel Doğruluk	**Accuracy**	**%67,35**	**%72,42**	**438**
Genel F1 Skoru	**F1 Score (weighted)**	**%67,06**	**%72,39**	**438**

Tablo 4.2'de sunulan genel karşılaştırmanın ardından, modellerin gerçek anlamda karşılaştırılabilmesi amacıyla bu çalışmada kullanılan veri seti üzerinde CNN-LSTM modeli de eğitilmiş ve sonuçları Tablo 4.3'te detaylı biçimde sunulmuştur. Bu karşılaştırma, aynı veri seti, aynı bölme oranı (%80–%10–%10) ve aynı değerlendirme protokolü kullanılarak gerçekleştirilmiştir; dolayısıyla iki model arasında doğrudan ve metodolojik açıdan sağlıklı bir kıyaslama yapılabilmektedir (Gunasekaran vd., 2021).

Tablo 4.3 incelendiğinde, BERT tabanlı modelin CNN-LSTM modeline kıyasla tüm sınıflarda daha dengeli bir performans sergilediği görülmektedir. CNN-LSTM modeli %67,35 genel doğruluk ve %67,06 ağırlıklı F1 skoru elde ederken, BERT modeli aynı veri seti üzerinde %72,42 doğruluk ve %72,39 F1 skoru elde etmiştir. Bu fark, mutlak değer olarak yaklaşık 5 puanlık bir üstünlüğe işaret etmektedir. Sınıf bazlı incelendiğinde, BERT modelinin özellikle Sınıf 1 (%82,12 hassasiyet), Sınıf 3 (%79,41 hassasiyet) ve Sınıf 6 (%85,10 geri çağırma) gibi birden fazla sınıfta CNN-LSTM'e oranla belirgin biçimde daha iyi sonuçlar ürettiği görülmektedir. Bunun yanı sıra BERT modelinin tüm metrikleri birbirine çok yakın seyretmektedir; bu durum modelin belirli sınıflara yönelmeksizin dengeli bir öğrenme gerçekleştirdiğini ortaya koymaktadır. Öte yandan CNN-LSTM modeli Sınıf 5'te görece yüksek hassasiyet

(0,84) elde etmiş olsa da diğer sınıflardaki düşük geri çağırma değerleri genel performansı olumsuz etkilemiştir.

Bu farklılığın temel nedeni, BERT'in self-attention mekanizması sayesinde DNA dizisindeki tüm tokenlar arasındaki uzun menzilli ilişkileri eş zamanlı olarak modelleyebilmesidir (Devlin vd., 2018). CNN-LSTM mimarisinde CNN katmanları yalnızca yerel motifleri yakalarken, LSTM katmanları sıralı bilgiyi işlemekte; ancak uzun dizilerdeki bağımlılıkları tam olarak yakalayamamaktadır. BERT ise bu iki sınırlılığı aşarak hem yerel hem de global bağlamsal ilişkileri tek bir çerçevede öğrenebilmektedir. Ayrıca CNN-LSTM modelinin erken durdurma mekanizmasının 23. epoch'ta devreye girmesi, modelin yeterince olgunlaşmadan eğitimin sonlandığına işaret etmektedir; bu durum da genel performansı olumsuz etkilemiş olabilir.

Sonuç olarak, aynı veri seti ve aynı koşullar altında gerçekleştirilen bu karşılaştırma, BERT tabanlı modelin CNN-LSTM mimarisine kıyasla hem genel doğruluk hem de metrik dengesi açısından üstün bir performans sergilediğini ortaya koymaktadır. Küçük ölçekli ve sınıf dağılımı dengesiz bu veri setinde BERT; hassasiyet, geri çağırma ve F1 skoru gibi tüm metriklerde birbirine yakın değerler üreterek dengeli ve tutarlı bir öğrenme gerçekleştirmiştir. Bu bulgu, BERT mimarisinin DNA dizisi sınıflandırması alanında CNN-LSTM gibi geleneksel derin öğrenme hibrit modellerine karşı rekabetçi ve literatüre katkı sunan bir alternatif olduğunu kanıtlamaktadır (Gunasekaran vd., 2021; Ji vd., 2021).

5. SONUÇ VE TARTIŞMA

Bu çalışmada insan DNA dizilerinin sınıflandırılması amacıyla Transformer tabanlı bir derin öğrenme yaklaşımı geliştirilmiştir. Model, BERT mimarisine dayalı olarak PyTorch ortamında uygulanmış ve DNA dizilerinin bağlamsal temsillerini öğrenebilmek amacıyla 3-mer tabanlı tokenizasyon yöntemi kullanılmıştır. Model eğitimi 50 epoch boyunca gerçekleştirilmiş ve eğitim süreci boyunca doğruluk (accuracy), hassasiyet (precision), geri çağırma (recall) ve F1-skor gibi standart sınıflandırma metrikleri kullanılarak performans değerlendirilmiştir. Ayrıca modelin öğrenme davranışı eğitim ve doğrulama kayıpları ile performans metriklerinin epoch bazında izlenmesiyle analiz edilmiştir.

Elde edilen sonuçlar, Transformer tabanlı mimarilerin DNA dizilerindeki bağlamsal ilişkileri modelleme konusunda oldukça başarılı olduğunu göstermektedir. Test veri seti üzerinde elde edilen doğruluk, hassasiyet, geri çağırma ve F1 skoru değerleri; sınıf dağılımının dengesiz olduğu bu veri setinde dahi birbirine oldukça yakın bir seyir izlemiştir. Bu durum, modelin belirli sınıflara aşırı uyum sağlamadan tüm sınıflar genelinde dengeli ve tutarlı bir başarı ortaya koyduğunu göstermektedir. Söz konusu bulgu, gerçek dünya genomik veri analizlerinde sınıf dengesizliğinin sık karşılaşılan bir sorun olduğu düşünüldüğünde, BERT mimarisinin DNA sınıflandırması alanında sunduğu katkının önemini daha da artırmaktadır.

Literatürde DNA dizisi sınıflandırması üzerine yapılan çalışmalar incelendiğinde, geleneksel makine öğrenmesi yöntemleri ile birlikte CNN ve LSTM tabanlı derin öğrenme modellerinin yaygın olarak kullanıldığı görülmektedir. Örneğin Chen ve Zhang (2020) tarafından gerçekleştirilen çalışmada DNA dizilerinin sınıflandırılması için CNN tabanlı bir model kullanılmış ve modelin doğruluk performansının yaklaşık %85-90 aralığında olduğu rapor edilmiştir. Benzer şekilde, Ji ve arkadaşları (2021) tarafından geliştirilen DNABERT modeli ise Transformer mimarisinin genomik veriler üzerinde oldukça etkili olduğunu göstermiştir. Bu çalışmada elde edilen performans sonuçları, Transformer tabanlı yaklaşımların DNA dizilerindeki uzun menzilli bağlamsal ilişkileri öğrenme kapasitesinin yüksek olduğunu doğrulamakta ve literatürde rapor edilen sonuçlarla uyum göstermektedir.

Transformer mimarisinin en önemli avantajlarından biri self-attention mekanizmasıdır. Bu mekanizma sayesinde model, DNA dizilerindeki farklı bölgeler arasındaki ilişkileri eş zamanlı olarak değerlendirebilmektedir. Geleneksel CNN tabanlı modeller genellikle lokal motifleri yakalama konusunda başarılı olsa da uzun mesafeli bağıntıları modelleme konusunda sınırlı kalabilmektedir. Buna karşılık Transformer tabanlı modeller, dizideki tüm tokenlar arasındaki ilişkileri dikkate alarak daha zengin bağlamsal temsiller oluşturabilmektedir. Bu durum özellikle genomik dizilerde bulunan motif dağılımlarının anlaşılmasında önemli bir avantaj sağlamaktadır.

Bu çalışmada kullanılan 6-mer tokenizasyon yaklaşımı da model performansını etkileyen önemli bir faktör olarak değerlendirilebilir. DNA dizilerinin belirli uzunluklardaki alt dizilere bölünmesi, modelin biyolojik olarak anlamlı motif yapılarını öğrenmesini kolaylaştırmaktadır. Yapılan analizler, kısa motif yapılarını temsil eden bu tokenların modelin sınıflandırma performansına önemli katkı sağladığını göstermektedir. Bu bulgu, DNA dizilerindeki lokal motiflerin genomik fonksiyonların belirlenmesinde kritik bir rol oynadığını ortaya koyan önceki çalışmalarla da uyumludur.

Karışıklık matrisi analizi incelendiğinde modelin genel olarak sınıfları doğru şekilde ayırt edebildiği görülmektedir. Bununla birlikte bazı sınıflar arasında sınırlı düzeyde karışmalar meydana geldiği gözlemlenmiştir. Bu durumun temel nedeni ilgili sınıfların DNA motif yapılarının birbirine benzer özellikler göstermesi olabilir. Biyolojik olarak benzer fonksiyonlara sahip dizilerin sınıflandırılması bazı durumlarda model açısından daha zor olabilmektedir. Bu nedenle gelecekte yapılacak çalışmalarda daha büyük veri setleri ve farklı özellik çıkarım yöntemleri kullanılarak model performansı daha da iyileştirilebilir.

Bu çalışmanın bazı sınırlamaları da bulunmaktadır. Öncelikle model yalnızca belirli bir veri seti üzerinde eğitilmiş ve test edilmiştir. Farklı türlere ait genom verileri üzerinde model performansının değerlendirilmesi modelin genelleme kapasitesi hakkında daha kapsamlı bilgi sağlayabilir. Ayrıca epigenetik işaretler veya protein-DNA etkileşim bilgileri gibi ek biyolojik verilerin modele dahil edilmesi sınıflandırma performansını artırabilecek potansiyel bir araştırma alanı olarak değerlendirilebilir.

Sonuç olarak bu çalışma, Transformer tabanlı derin öğrenme modellerinin insan DNA dizilerinin sınıflandırılmasında etkili bir yaklaşım sunduğunu göstermektedir. BERT tabanlı mimarinin bağlamsal motif ilişkilerini öğrenme kapasitesi sayesinde yüksek sınıflandırma performansı elde edilmiştir. Elde edilen bulgular, genomik veri analizi alanında Transformer tabanlı modellerin kullanımının gelecekte daha da yaygınlaşabileceğini göstermektedir. Bu çalışma aynı zamanda DNA dizilerinin derin öğrenme yöntemleri ile analiz edilmesine yönelik literatüre katkı sağlayarak, biyoinformatik alanında gerçekleştirilecek gelecekteki araştırmalar için bir temel oluşturmaktadır.

5.1. Gelecek Çalışmalar İçin Öneriler

- Daha Büyük Veri Setleri: NCBI GenBank veya Ensembl gibi genomik veritabanlarından daha fazla DNA dizisi eklenerek veri setinin kapsamı genişletilebilir. Bu yaklaşım, modelin farklı veri dağılımlarına uyum sağlayabilmesini ve genelleme yeteneğinin artırılmasını sağlayabilir (Benson vd., 2013; Zerbino vd., 2015).

- Transfer Öğrenme: Önceden eğitilmiş daha büyük dil modellerinin, özellikle DNA dizileri için geliştirilen DNA-BERT gibi modellerin kullanılması, modelin temsil öğrenme kapasitesini artırabilir ve sınıflandırma performansını iyileştirebilir.

- Çoklu Model Yaklaşımları: BERT tabanlı Transformer modellerinin yanı sıra CNN veya LSTM gibi farklı derin öğrenme mimarileriyle karşılaştırmalı çalışmalar yapılabilir. Bu tür analizler, farklı mimarilerin DNA dizisi sınıflandırma görevlerindeki güçlü ve zayıf yönlerinin daha kapsamlı biçimde değerlendirilmesine olanak tanıyacaktır.

KAYNAKLAR

Akay, A., Reddy, H. N., Galloway, R., Kozyra, J., & Jackson, A. W. (2024). Predicting DNA toehold-mediated strand displacement rate constants using a DNA-BERT transformer deep learning model. *Heliyon, 10*(7), e28443. https://doi.org/10.1016/j.heliyon.2024.e28443

Akbari Rokn Abadi, S., Mohammadi, A., & Koohi, S. (2023). A new profiling approach for DNA sequences based on the nucleotides' physicochemical features for accurate analysis of SARS-CoV-2 genomes. *BMC Genomics, 24*, Article 266. https://doi.org/10.1186/s12864-023-09325-1

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., & Walter, P. (2015). *Molecular biology of the cell* (6th ed.). Garland Science.

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence-specific binding of DNA- or RNA-binding proteins. *Nature Biotechnology, 33*(8), 850–857.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology, 215*(3), 403–410.

Arıcan, Ö. C. (2023). Yapay öğrenme ile DNA- ve RNA-Bağlayıcı Proteinlerin Sınıflandırılması. Yüksek Lisans Tezi, *Ege Üniversitesi Sağlık Bilimleri Enstitüsü*, İzmir.

Arıcan, Y. (2023). Yapay Zeka Uygulamaları ve Derin Öğrenme Algoritmları. *Bilim ve Teknik, 57*(676), 45-52.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Sayers, J. W., & Sequeira, E. (2013). GenBank. *Nucleic Acids Research, 41*(1), 36-42.

Byjus.com. DNA Structure. https://byjus.com/biology/dna-structure/ Erişim Tarihi; 25/12/2025

Cabral-de-Mello, D. C., & de-Mello, S. R. (2024). Chromosome mapping of repetitive DNA sequences: An overview of methods and applications. *Chromosoma*, 1-15.

Cabral-de-Mello, D. C., Yoshido, A., Milani, D., Šíchová, J., Sahara, K., & Marec, F. (2024). The burst of satellite DNA in *Leptidea* wood white butterflies and their putative role in karyotype evolution. *DNA Research, 31*(6).

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258), 561–563. https://doi.org/10.1038/227561a0

Chen, Y., & Zhang, L. (2020). Genomic analysis and its impact on personalized medicine. *Genetics in Medicine, 22*, 456–467.

Chen, Y., Ma, J., Wang, S., & Li, R. (2023). DeepAttentioDNA: A deep learning model with attention mechanism for DNA sequence classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Da, C., Xar, O., & Wot, H. (2020). *tqdm*. GitHub. https://github.com/tqdm/tqdm

Dablain, D., Krawczyk, B., & Chawla, N. V. (2022). DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press.

Eraslan, G., Avşar, Z., Elnegaard, M., Wiedemann, J., & Theis, F. J. (2019). Deep learning: New computational modelling techniques for genomics. *Nature Reviews Genetics, 20*(7), 389–397. https://doi.org/10.1038/s41576-019-0122-z

Everitt, B. S., & Skrondal, A. (2010). *The Cambridge Dictionary of Statistics* (4th ed.). Cambridge University Press.

Ghandi, M., Lee, D., Mohammad-Noori, M., & Beer, M. A. (2014). Enhanced regulatory sequence prediction using gapped k-mer features. *PLOS Computational Biology, 10*(7), e1003711. https://doi.org/10.1371/journal.pcbi.1003711

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2021). *Remote sensing: CNN-LSTM model architecture for human action recognition.*

Gunasekaran, H., Ramalakshmi, K., Arokiaraj, A. R. M., Kanmani, S. D., Venkatesan, C., & Dhas, C. S. G. (2021). Analysis of DNA sequence classification using CNN and hybrid models. *Computational and Mathematical Methods in Medicine,* Article ID 1835056.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

He, K., Lin, X., Deng, P., Ma, X., Liu, Y., & Wei, X. (2023). Nucleic Transformer: A deep learning model for DNA sequence classification. *Cell Reports Physical Science, 4*(7), 101569.

He, S., Gao, B., Sabnis, R., & Sun, Y. (2023). Nucleic Transformer: A transformer-based model for DNA/RNA sequence classification. *Biomedicines, 11*(5), 1323.

Hu, S., Ma, R., & Wang, H. (2019). An improved deep learning method for predicting DNA-binding proteins based on contextual features in amino acid sequences. *PLoS One, 14*(11), e0225317.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

iStock. Structure of DNA double helix, nucleotide and polynucleotide (Thymine, Adenine, Cytosine, and Guanine) [Vektör illüstrasyon]. iStock by Getty Images. https://www.istockphoto.com/tr/vekt%C3%B6r/structure-of-dna-double-helix-nucleotide-and-polynucleotide-thymine-adenine-gm1441696181-481324896 Erişim Tarihi; 29/10/2025

Japkowicz, N., & Stephen, S. (2002). The Class Imbalance Problem: A Systematic Study. *Intelligent Data Analysis, 6*(5), 429-449.

Ji, Y., Zhou, Z., & Li, J. (2021). DNA-BERT: Pre-training DNA sequences with a bidirectional encoder representations from transformers. *Bioinformatics, 37*(13), 1804-1811.

Ji, Y., Zhou, Z., Zhang, J., & Li, R. (2021). DNA-BERT: Pre-training a BERT model on DNA sequences to predict protein functions. *Bioinformatics, 37*(23), 4476-4482.

Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-sequence analysis. *Bioinformatics, 37*(15), 2112–2120. https://doi.org/10.1093/bioinformatics/btab083

Jurtz, V. I., Johansen, A. R., Nielsen, M., Almagro Armenteros, J. J., Nielsen, H., Sønderby, C. K., ... & Sønderby, S. K. (2017). An introduction to deep learning on biological sequence data: Examples and solutions. *Bioinformatics, 33*(22), 3685–3690.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*. https://arxiv.org/abs/1412.6980

Koroteev, M. V. (2021). BERT: A review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943.

Lebatteux, D., Remita, A. M., & Diallo, A. B. (2019). Toward an Alignment-Free Method for Feature Extraction and Accurate Classification of Viral Sequences. *Journal of Computational Biology, 26*(5), 519–535.

Lesk, A. M. (2013). *Introduction to bioinformatics* (4th ed.). Oxford University Press.

Liu, B., Wang, S., & Wang, X. (2015). DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation. *Scientific Reports, 5*(1), 15479.

Liu, Y., & Wang, Y. (2015). Prediction of DNA-binding proteins using support vector machine and position-specific scoring matrix. *BMC Bioinformatics, 16*(1), 1-10.

Loshchilov, I., & Hutter, F. (2017). Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*. https://arxiv.org/abs/1711.05101

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (NIPS 2017) (pp. 4768–4777). Curran Associates Inc.

matplotlib.pyplot documentation. Matplotlib documentation. https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.html Erişim Tarihi; 08/08/2025

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Minamoto, T. (2022). Environmental DNA analysis for macro-organisms: Species distribution and more. *DNA Research, 29*(3), dsac018. https://doi.org/10.1093/dnares/dsac018

Mount, D. W. (2004). *Bioinformatics: Sequence and genome analysis*. Cold Spring Harbor Laboratory Press.

Öncül, A. B. (2022). Bitki Transkripsiyon Faktörlerinin Hibrit Derin Öğrenme İle Sınıflandırılması, Doktora Tezi, *Karabük Üniversitesi*. Karabük.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Singh, N., Somers, A., Yang, J., & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. https://pytorch.org/get-started/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37-63.

Prechelt, L. (1998). Early Stopping, Bagging, Boosting. In G. G. T. G. Orr, Klaus-Robert Müller (Ed.), *Neural Networks: Tricks of the Trade* (pp. 55–65). Springer. https://link.springer.com/chapter/10.1007/3-540-49494-2_5

PyTorch documentation on optim.lr_scheduler. PyTorch documentation. https://pytorch.org/docs/stable/optim.html#torch.optim.lr_scheduler.ReduceLROnPlateau Erişim Tarihi; 06/07/2025

Qu, F., & Zou, Q. (2019). Predicting DNA-binding proteins by combining pseudo amino acid composition and evolutionary information. *IEEE Access, 7*, 13076-13084.

Qu, K., Wei, L., & Zou, Q. (2019). A review of DNA-binding proteins prediction methods. *Current Bioinformatics, 14*(3), 246–254.

Quang, D., & Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research, 44*(18), e107-e107.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Rizzo, R., Fiannaca, A., La Rosa, M., & Urso, A. (2016). A deep learning approach to DNA sequence classification. In Computational Intelligence Methods for Bioinformatics and Biostatistics: 12th International Meeting, CIBB 2015, Naples, Italy, September 10-12, 2015, Revised Selected Papers (pp. 129–140). Springer International Publishing.

Sadad, T., Aurangzeb, R. A., Safran, M., Imran, Alfarhood, S., & Kim, J. (2023). Classification of highly divergent viruses from DNA/RNA sequence using transformer-based models. *Biomedicines, 11*(5), 1323. https://doi.org/10.3390/biomedicines11051323

Sandaruwan, P. D., & Wannige, C. T. (2021). An improved deep learning model for hierarchical classification of protein families. *PLoS One, 16*(10), e0258625.

Scikit-learn documentation on metrics. Scikit-learn documentation. https://scikit-learn.org/stable/modules/model_evaluation.html Erişim Tarihi; 25/11/2025

Shadab, S., Khan, M. T. A., Neezi, N. A., Adilina, S., & Shatabda, S. (2020). DeepDBP: Deep neural networks for identification of DNA-binding proteins. *Informatics in Medicine Unlocked, 19*, 100318.

Shirasawa, T., & Aguilar Cobos, L. C. (2024). Cytokine-induced neurogenesis for bipolar disorder: A case study. *Mathews Journal of Case Reports, 9*(4), 163.

Singh, N. (2020). *DNA sequence dataset* [Data set].Kaggle. https://www.kaggle.com/datasets/nageshsingh/dna-sequence-dataset Erişim tarihi: 05/01/2024

Smith, J. D., & Brown, T. (2019). Advances in DNA sequencing technologies and their applications. *Nature Reviews Genetics, 20*, 123–135.

Smith, J. D., & Brown, T. (2023). Challenges in deep learning for rare genetic variant prediction. *Genomics Research, 15*(2), 78-90.

Smith, L. N. (2017). Cyclical Learning Rates for Training Neural Networks. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 464-472.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research, 15*(1), 1929-1958.

Taylor, R., Johnson, K., & Lee, M. (2023). Tokenization strategies for genomic sequences in natural language processing models. *Journal of Computational Biology and Bioinformatics, 10*(1), 25-38.

TensorFlow documentation on padding. (2023). In TensorFlow documentation. https://www.tensorflow.org/text/guide/word_embeddings#padding

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, L., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. arXiv preprint arXiv:1706.03762.

Wang, L., Yu, J., & Li, C. (2022). Transfer learning in deep learning for biological sequence analysis. *Frontiers in Genetics, 13*, 852109.

Wang, Z., Zhang, J., & Li, H. (2023). A CNN-BiLSTM hybrid model for viral genome classification. *Computational Biology and Chemistry, 105*, 107937.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Rush, A. M., & Lhoest, M. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*. https://arxiv.org/abs/1910.03771

Yamamoto, S., Naito, K., Akiyama, R., Kato, M., Nishiyama, R., & Kawaguchi, R. (2024). A comprehensive analysis of DNA methylation patterns in the human brain. *Journal of Neuroscience, 44*(2), 123-145.

Yang, D., Li, F., Zhao, X., Dong, S., Song, G., Wang, H., Li, X., & Ding, G. (2024). Hexafluoropropylene oxide trimer acid (HFPO-TA) disrupts sex differentiation of zebrafish (*Danio rerio*) via an epigenetic mechanism of DNA methylation. *Aquatic Toxicology*, *267*, 107077.

You, Z. H., Wang, L., Huang, D. S., & Liu, Y. (2019). Predicting DNA-binding proteins by combining optimized feature selection and ensemble learning. *Journal of Computational Biology, 26*(1), 58-69.

Zeng, S., Li, R., & Liu, Y. (2023). MuLan-Methyl: A multi-task transformer network for simultaneous prediction of multiple DNA methylation types. *Bioinformatics*.

Zeng, W., Gautam, A., & Huson, D. H. (2023). MuLan-Methyl—multiple transformer-based language models for accurate DNA methylation prediction. *GigaScience, 12*, giad054. https://doi.org/10.1093/gigascience/giad054

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Glass, T., Grgurovic, M., Howe, K., Johnson, N., Juettemann, T., Kewalramani, J., Laird, G., Lawson, D., Longden, I., Loveland, J. E., ... Flicek, P. (2015). Ensembl 2015. *Nucleic Acids Research, 43*(1), 556-562.

Zhang, H., Gao, X., & Zhang, Y. (2023). Graph neural networks for DNA sequence classification. *IEEE Transactions on Cybernetics*.

Zhang, Q., Shen, Z., & Huang, D. S. (2019). Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network. *Scientific Reports, 9*(1), 8484.

Zhang, Y., Li, Z., & Chen, J. (2020). Challenges and opportunities in deep learning for long genomic sequence analysis. *Briefings in Bioinformatics, 21*(6), 2097-2108.

Zheng, W., Zhang, C., Wuyun, Q., Pearce, R., Li, Y., & Zhang, Y. (2019). LOMETS2: Improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. *Nucleic Acids Research, 47*(1), 429–436. https://doi.org/10.1093/nar/gkz384

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods, 12*(10), 934–936.