

**T.C.
ERCIYES ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

**METAGENOM VERİSİNDEN KRONİK HASTALIKLARA
AİT KARARLI BİYOBELİRTEÇLERİN TESPİTİNE
YÖNELİK ALGORİTMALARIN GELİŞTİRİLMESİ**

**Hazırlayan
Umay Gülfem ELGÜN ÇİFTÇİOĞLU**

**Danışman
Doç. Dr. Özkan Ufuk NALBANTOĞLU**

Doktora Tezi

**Aralık 2024
KAYSERİ**

**T.C.
ERCIYES ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

**METAGENOM VERİSİNDEN KRONİK HASTALIKLARA
AİT KARARLI BİYOBELİRTEÇLERİN TESPİTİNE
YÖNELİK ALGORİTMALARIN GELİŞTİRİLMESİ
(Doktora Tezi)**

**Hazırlayan
Umay Gülfem ELGÜN ÇİFTÇİOĞLU**

**Danışman
Doç. Dr. Özkan Ufuk NALBANTOĞLU**

**Bu çalışma TÜBİTAK tarafından 2211/C programı kapsamında
desteklenmiştir.**

**Aralık 2024
KAYSERİ**

BİLİMSEL ETİĞE UYGUNLUK

Bu çalışmadaki tüm bilgilerin, akademik ve etik kurallara uygun bir şekilde elde edildiğini beyan ederim. Aynı zamanda bu kural ve davranışların gerektirdiği gibi, bu çalışmanın özünde olmayan tüm materyal ve sonuçları tam olarak aktardığımı ve referans gösterdiğimi belirtirim.

Umay Gülfem ELGÜN ÇİFTÇİOĞLU

İmza:

“Metagenom Verisinden Kronik Hastalıklara Ait Kararlı Biyobelirteçlerin Tespitine Yönelik Algoritmaların Geliştirilmesi” adlı Doktora tezi, Erciyes Üniversitesi Lisansüstü Tez Önerisi ve Tez Yazma Yönergesi’ne uygun olarak hazırlanmıştır.

Hazırlayan

Umay Gülfem ELGÜN ÇİFTÇİOĞLU

Danışman

Doç. Dr. Özkan Ufuk NALBANTOĞLU



Bilgisayar Mühendisliği ABD Başkanı

Prof. Dr. Veysel ASLANTAŞ

TEŞEKKÜR

Çalışmalarında deneyimleriyle bana rehberlik eden, bilgeliği ve çalışma disiplini ile daima en iyiye hedeflemem için teşvik eden, araştırma sürecimde her zaman güvenilir bir destek ve ilham kaynağı olan, engin bilgisine rağmen tevazu dolu yaklaşımı ve insan ilişkileri ile hem akademik hem sosyal yaşamda örnek aldığım kıymetli danışman hocam Doç. Dr. Özkan Ufuk Nalbantoğlu'na sonsuz teşekkürlerimi sunarım. Tez izleme süreci boyunca değerli fikirleri ve geri bildirimleri ile tez çalışmasına katkı sağlayan saygıdeğer hocalarım Prof. Dr. Aycan Gündoğdu ve Prof. Dr. Celal Öztürk'e teşekkür ederim. Çalışmalarım esnasında desteklerini esirgemeyen, Meryem Altın Karagöz, Mehmet Hora, Samet Saka ve tüm çalışma arkadaşlarıma teşekkür ederim.

Tez çalışmamı 2211-C Yurt İçi Öncelikli Alanlar Doktora Burs Programı ile destekleyen TÜBİTAK Bilim İnsanı Destekleme Daire Başkanlığı birimine teşekkür ederim.

Son olarak, tez süresince her türlü yardımı, anlayışı, motivasyonu ve desteği sağlayan yol arkadaşım, sevgili eşim Burak Çiftcioğlu'na ve başta çalışmam için gerekli ortamı her koşulda sağlayan, ilgi ve sevgisini her zaman hissettiren kıymetli annem Hatice Elgün olmak üzere, desteğini esirgemeyen aileme teşekkürlerimi sunarım.

METAGENOM VERİSİNDEN KRONİK HASTALIKLARA AİT KARARLI BİYOBELİRTEÇLERİN TESPİTİNE YÖNELİK ALGORİTMALARIN GELİŞTİRİLMESİ

Umay Gülfem ELGÜN ÇİFTÇİOĞLU

**Erciyes Üniversitesi, Fen Bilimleri Enstitüsü
Doktora Tezi, Kasım 2024**

Danışman : Doç. Dr. Özkan Ufuk NALBANTOĞLU

ÖZET

Bağırsak mikrobiyomu, hastalık tanısı ve prognozu için hayati bilgiler sağlayarak insan sağlığında önemli bir rol oynamaktadır. Metagenomik uygulama sayısındaki artış, yeni nesil dizileme maliyetlerindeki düşüş ve mikrobiyom işlevlerinin derinlemesine anlaşılma isteği, genetik veri üretiminde hızlı bir artışa yol açmaktadır. Mikrobiyom analizi için, genellikle kohort başına, yüzlerce veya çoğunlukla binlerce veri üretilmektedir. Bu süreçte pratik ve teknik sınırlamalar olduğu için, az örnek sayısı-yüksek boyut sorunu, hipotezlerin testinde çeşitli zorlukları beraberinde getirmektedir. Büyük miktarlarda biyolojik verinin analizi için, gelişmiş matematiksel ve istatistiksel teknikleri kullanmak önemlidir. Son yıllarda, metagenomik veri analizinin merkezinde yer alan istatistik, derin öğrenme (DL) ve veri madenciliği alanları önemli ölçüde ilerleme kaydetmiştir. Yapılan çalışmalar ile DL' nin hacimli verilerden yeni biyolojik hipotezler veya keşifler elde edilebileceği ve ilgili metagenom verilerindeki kalıpları ayırt edebilme ve tanımlama potansiyeli gösterilmiştir. Bu tez çalışmasında bağırsak mikrobiyomu verilerindeki yüksek boyutluluk, veri standardizasyonu ve sınırlı verilerle ilgili zorluklara çözüm bulmak amacıyla yeni algoritmalar geliştirilmiştir. Bu algoritmalar, yüksek örneğe sahip birleştirilmiş hastalık verileri ile test edilmiş ve fenotip tahmini, kararlı biyobelirteç keşfi, tutarlılık analizi ve kıt verilerde sınıflandırma gibi birçok görevde üstün performans göstermiştir. Geliştirilen yaklaşımlar, genellenebilir olması, sağlamlığı, uyarlanabilirliği, genişletilebilir ve taşınabilir olması sebebiyle bir sağlık indeksi olarak hizmet edebilir ve klinik uygulamalar için umut verici bir yön sunabilir.

Anahtar Kelimeler: Derin Öğrenme, Mikrobiyom Analizi, Taksonomik Sınıflandırma, Transfer Öğrenme, Evrişimli Sinir Ağları, Biyobelirteç Keşfi

DEVELOPMENT OF ALGORITHMS FOR THE DETECTION OF ROBUST BIOMARKERS ON CHRONIC DISEASES FROM METAGENOME DATA**Umay Gülfem ELGÜN ÇİFTÇİOĞLU****Erciyes University, Graduate School of Natural and Applied Sciences
PhD Thesis, Nov 2024****Supervisor: Assoc. Prof. Özkan Ufuk NALBANTOĞLU****ABSTRACT**

The gut microbiome plays an important role in human health by providing vital information for disease diagnosis and prognosis. The increase in the number of metagenomics applications, the decrease in the costs of next generation sequencing, and desire for an in-depth understanding of microbiome functions have led to a rapid increase in genetic data production. The metagenome data has typically been produced in the orders of tens, hundreds or mostly in thousands per cohort. As there are practical and technical limitations resulting in this process, the small sample size-large dimension problem brings along hypothesis testing difficulties. It is important to use advanced mathematical and statistical techniques for the analysis of biological data. In recent years, the fields of statistics, deep learning (DL), and data mining have made significant progress. Studies have shown that DL has the potential to distinguish and identify patterns in metagenome data. In this thesis, new algorithms have been developed to find solutions to the challenges of high dimensionality, data standardization, and limited data in the gut microbiome datasets. These algorithms have been tested with large sample size merged data and have shown superior performance in many tasks such as phenotype prediction, robust biomarker discovery, consistency analysis, and classification with scarce data. Proposed approaches can serve as a wellness index due to their generalizability, robustness, adaptability, extensibility, and portability, and may offer a promising direction for clinical applications.

Keywords: Deep Learning, Microbiome Analysis, Taxonomic Classification, Transfer Learning, Convolutional Neural Networks, Biomarker Discovery

İÇİNDEKİLER

METAGENOM VERİSİNDEN KRONİK HASTALIKLARA AİT KARARLI BİYOBELİRTEÇLERİN TESPİTİNE YÖNELİK ALGORİTMALARIN GELİŞTİRİLMESİ

BİLİMSEL ETİĞE UYGUNLUK SAYFASI	ii
YÖNERGEYE UYGUNLUK SAYFASI	iii
KABUL VE ONAY	iv
TEŞEKKÜR	v
ÖZET	vi
ABSTRACT	vii
İÇİNDEKİLER	viii
KISALTMALAR	xii
TABLolar LİSTESİ	xiv
ŞEKİLLER LİSTESİ	xv
GİRİŞ	1

1. BÖLÜM

GENEL BİLGİLER VE LİTERATÜR ÇALIŞMASI

1.1. Mikrobiyomun Sağlık ve Hastalık ile İlişkisi	8
1.2. Mikrobiyom Veri Analizinde Kullanılan Yöntemler	11
1.2.1. Mikrobiyom ve Makine Öğrenmesi	11
1.2.1.1. Klasik Makine Öğrenmesi Yöntemleri	13
1.2.1.2. Boyut İndirgeme için Kullanılan Makine Öğrenmesi Yöntemleri	16
1.2.1.3. Biyobelirteç Keşfi için Kullanılan Makine Öğrenmesi Yöntemleri	17
1.2.2. Mikrobiyom ve Derin Öğrenme	19
1.2.2.1. Mikrobiyom Verilerinde Kullanılan Derin Öğrenme Yöntemleri	20
1.3. Derin Öğrenme ile Mikrobiyom Analizinde Karşılaşılan Problemler . .	22

1.4. Araştırmanın Amacı ve Önemi	25
--	----

2. BÖLÜM

ARAŞTIRMA YÖNTEMLERİ

2.1. Gürbüz Temel Bileşen Analizi	28
2.2. Yapay Sinir Ağları	30
2.2.1. Aktivasyon Fonksiyonları	32
2.2.2. Kayıp Fonksiyonları	34
2.2.3. Geri Yayılım	35
2.2.4. Optimizasyon Algoritmaları	37
2.3. Derin Öğrenme Modelleri	38
2.3.1. Otokodlayıcı (Autoencoder)	38
2.3.1.1. Evrişimli Otokodlayıcı	41
2.3.2. U-Net	42
2.3.3. Üretici Çekişmeli Ağlar	43
2.4. Sınıflandırıcı Mimarileri	46
2.4.1. Evrişimli Sinir Ağları	46
2.5. Transfer Öğrenme	48
2.6. Performans Değerlendirme Metrikleri	49
2.7. Mikrobiyom Verilerinde Sınıflandırma Analizi için Kullanılan Derin Öğrenme Tabanlı Çalışmalar	53
2.8. Mikrobiyom Verilerinde Biyobelirteç Keşfi için Kullanılan Derin Öğrenme Tabanlı Çalışmalar	56

3. BÖLÜM

MİKROBİYOM VERİLERİNDE SINIFLANDIRMA PERFORMANSININ GELİŞTİRİLMESİ

3.1. DeepGum: Mikrobiyom Analizi için Derin Öznitelik Aktarımı	58
3.1.1. Genel Bakış	58
3.1.2. Çalışmada Kullanılan Mikrobiyom Veri Kümeleri ve Ön İşleme	58
3.1.3. Denetimsiz Yöntemlerle Derin Öznitelik Çıkarımı	59
3.1.4. Örneklem Büyüklüğüne Dayalı Sınıflandırma Analizi	64

3.1.5. DeepGum Yöntemiyle Elde Edilen Sınıflandırma Analizi Sonuçları	64
3.1.5.1. Bağırsak Hastalıkları Veri Kümelerinde DeepGum ile Elde Edilen Sınıflandırma Sonuçları	64
3.1.5.2. Rastgele Seçilmiş Seyrek Örnekli Kohortlar için Sınıflandırma Analizi Sonuçları	67
3.2. Deeptaxim: Taksonomik Veri Kümeleri için Görüntü Tabanlı Derin Öğrenme Modelleri ile Kapsamlı Sınıflandırma Analizi	68
3.2.1. Genel Bakış	68
3.2.2. Kullanılan Veri Kümelerinin Detayları ve Ön İşleme	69
3.2.2.1. Mikrobiyom Verilerinden Görüntü Verilerine Geçiş Süreci	70
3.2.3. Çalışmada Kullanılan Denetimsiz Derin Öğrenme Mimarileri	70
3.2.4. Deeptaxim Modelinin Mikrobiyom-görüntü Veri Kümelerine Uygulanması	73
3.2.5. Deeptaxim Yöntemiyle Elde Edilen Sınıflandırma Analizi Sonuçları	75
3.2.5.1. Gastro Veri Kümesinde Deeptaxim ve Mevcut Yöntemlerin Sınıflandırma Performansının Karşılaştırılması	75
3.2.5.2. Çoklu Hastalık Veri Kümesinde Deeptaxim ve Mevcut Yöntemlerin Sınıflandırma Performansının Karşılaştırmalı Analizleri	76
3.2.5.2.1. Deeptaxim ile Sağlık-Hastalık (Wellness) İkili Sınıflandırma Analizi	76
3.2.5.2.2. İkili Sınıflandırmanın Hastalığa Özgü Fenotip Analizi	77
3.2.5.2.3. Bir Hastalığı Dışarıda Bırakma Yaklaşımının Değerlendirilmesi	78

4. BÖLÜM

MİKROBİYOM VERİLERİNDE BİYOBELİRTEÇ KEŞFİ

4.1. Disrobiom: Mikrobiyom Verileri için Derin Öğrenme Algoritmalarıyla Gürbüz Biyobelirteç Keşfi	80
4.1.1. Genel Bakış	80
4.1.2. Kullanılan Veri Kümeleri ve Ön İşlemler	80
4.1.3. Denetimsiz Derin Modeller	81
4.1.4. Derin Modellerle Öznitelik Seçimi	83
4.1.5. Disrobiom Yöntemiyle Elde Edilen Değerlendirme Sonuçları	86

4.1.5.1.	Disrobiom Sınıflandırma ve Tutarlılık Performansının Mevcut Yöntemlerle Karşılaştırılması	86
4.1.5.2.	Keşfedilen Potansiyel Biyobelirteçler için Çapraz Kohort Doğrulaması	89
4.1.5.3.	Keşfedilen Öznitelik Kümelerinin Farklı Veri Kümeleri için Tutarlılık Analizi	91
4.1.5.4.	İBH ve KK için Keşfedilen Potansiyel Biyobelirteçler	91

5. BÖLÜM

TARTIŞMA, SONUÇ ve ÖNERİLER

5.1. Tartışma	95
5.2. Sonuç ve Öneriler	98
KAYNAKÇA	102
ÖZGEÇMİŞ	122

KISALTMALAR

ACC	: Doğruluk
ACVD	: Aterosklerotik Kardiyovasküler Hastalık
AUC	: ROC Eğrisi Altındaki Alan
AS	: Ankilozan Spondilit
CAE	: Evrişimli Otokodlayıcı
CH	: Crohn Hastalığı
CNN	: Evrişimli Sinir Ağı
CV	: Çapraz Doğrulama
DL	: Deep Learning
DSA	: Derin Sinir Ağları
ENet	: Elastik Ağlar
FCNN	: Tam Bağlantılı Sinir Ağı
FFN	: İleri Beslemeli Sinir Ağı
FOBT	: Dışkıda Gizli Kan Testi
FS	: Feature Selection
GI	: Gastrointestinal
GAN	: Üretici Çekişmeli Ağlar
GB	: Gradient-Boosted
GH	: Graves Hastalığı
HBS	: Huzursuz Bağırsak Sendromu
HM	: Hata Matrisi
İBH	: İltihaplı Bağırsak Hastalığı
KI	: Kuncheva İndeksi
KK	: Kolorektal Kanser
KS	: Karaciğer Sirozu
Lodo	: Leave-one-disease-out
LOO	: Leave One Out
LR	: Lojistik Regresyon
MAE	: Mean Absolute Error
ML	: Machine Learning

MLP	: Çok Katmanlı Algılayıcı
mRMR	: Minimum Fazlalık- Maksimum Alaka
MS	: Multipl Skleroz
MSE	: Ortalama Kareysel Hata
NGS	: Yeni nesil dizileme
NAYKH	: Non-Alkolik Yağlı Karaciğer Hastalığı
NLP	: Natural Language Processing
NN	: Sinir Ağları
OTU	: Operasyonel Taksonomik Birim
PCA	: Temel Bileşen Analizi
PCoA	: Temel Koordinat Analizi
PCP	: Principal Component Pursuit
RA	: Romatoid Artrit
RF	: Rastgele Orman
RNN	: Yinelemeli Sinir Ağları
RPCA	: Robust Principal Component Analysis
rRNA	: ribozomal RNA
RT-PCR	: Gerçek Zamanlı Polimeraz Zincir Reaksiyonu
ROC	: Receiver Operating Characteristic
SGD	: Stokastik Gradyan İnişi
SVM	: Support Vector Machine
XGB	: Xtreme Gradient Boosting
T2D	: Tıp 2 Diyabet
ÜK	: Ülseratif Kolit
YSA	: Yapay Sinir Ağları

TABLOLAR LİSTESİ

Tablo 1.1.	Farklı hastalıklar için bağırsak mikrobiyomunda gözlemlenen değişimler.	10
Tablo 3.1.	Bağırsak mikrobiyomu veri kümelerine ait detaylı bilgiler.	59
Tablo 3.2.	Çoklu-Hastalık veri kümesine ait detaylı bilgiler.	69
Tablo 3.3.	Deeptaxim' in GI hastalıklar üzerindeki sınıflandırma performansının mevcut yöntemlerle (RF, XGB, MLP ve iMic) karşılaştırılması. . . .	76
Tablo 3.4.	Deeptaxim'in Çoklu hastalık veri kümesi üzerindeki sınıflandırma performansının mevcut yöntemlerle karşılaştırılması.	77
Tablo 3.5.	Sağlık değerlendirmesinin her model için hastalık başına doğru tahmin yüzdesi.	78
Tablo 4.1.	Veri Kümelerine ait detaylı bilgiler.	81
Tablo 4.2.	Metagenomik veri kümeleri üzerinde RPCA, U-Net ve AUTO modelleri ile seçilen 10, 20, 50 ve 100 öznitelik için karşılaştırmalı RF Sınıflandırma Performansı (AUC ve F1 puanı).	87

ŞEKİLLER LİSTESİ

Şekil 1.1.	Dizileme yöntemlerinin süreç adımlarını içeren iş akış şeması.	7
Şekil 1.2.	Mikrobiyom verilerinden hastalık teşhisi için model geliştirmenin ilk adımı olan veri üretimine yönelik basitleştirilmiş bir iş akışı.	13
Şekil 1.3.	ML/DL yöntemlerine dair kapsamlı süreci içeren genel bir akış şeması.	20
Şekil 2.1.	a: Biyolojik nöron ve sinir ağının temsili gösterimi, b: Temel yapay nöron gösterimi, c: Örnek bir yapay sinir ağı.	31
Şekil 2.2.	Sıklıkla kullanılan aktivasyon fonksiyonları ve matematiksel formülleri.	33
Şekil 2.3.	Temel bir otokodlayıcı mimarisi.	40
Şekil 2.4.	Temel Evrişimli Otokodlayıcı (CAE) mimarisi.	41
Şekil 2.5.	Temel U-Net mimarisi.	42
Şekil 2.6.	Temel Üretici Çekişmeli Ağ (GAN) mimarisi.	44
Şekil 2.7.	Temel CNN mimarisi.	47
Şekil 3.1.	DeepGum yönteminin genel mimarisi.	60
Şekil 3.2.	DeepGum katmanlarının detaylı bilgileri.	62
Şekil 3.3.	DeepGum'ın GI hastalıklar için tahmin performansı.	65
Şekil 3.4.	U-Net ve Otokodlayıcı modellerinin yeniden yapılandırılmış ve kodlayıcı modelleri için tahmin performansının karşılaştırılması.	66
Şekil 3.5.	Her veri setinden rastgele seçilen 5, 10, 20, 50 ve 100 örnek için, DeepGum+RF ile belirlenen ve RF sınıflandırıcı model ile belirlenen ortalama AUC puanlarının karşılaştırmalı sonuçları.	67
Şekil 3.6.	Deeptaxim'de kullanılan derin öğrenme modellerinin detaylı katman bilgileri.	72
Şekil 3.7.	Deeptaxim sürecinin genel görünümü.	74
Şekil 3.8.	Çoklu hastalık veri kümesinde, bir hastalığı dışarıda bırakma (Lodo) yaklaşımının değerlendirilmesi.	79

Şekil 4.1.	Derin modellerin detaylı katman bilgileri.	82
Şekil 4.2.	Modellerin girdi ve çıktıları arasındaki fark kullanılarak öznitelik seçme prosedürü.	84
Şekil 4.3.	Derin modellerin Öznitelik Seçimi ve Sınıflandırma-Tutarlılık performans değerlendirme çerçevesi.	85
Şekil 4.4.	Metagenomik veri kümeleri üzerinde RPCA, U-Net ve Otokodlayıcı (AUTO) modelleri ile seçilen 10, 20, 50 ve 100 öznitelik için karşılaştırmalı RF Sınıflandırma Performansı (ACC).	88
Şekil 4.5.	Metagenomik veri kümelerinde FS yöntemleri için tutarlılık performansı karşılaştırması.	88
Şekil 4.6.	Bir veri kümesi için en önemli 10 özneliğin, diğer veri kümeleri üzerindeki Otokodlayıcı, U-Net veya RPCA modelleri ile sınıflandırma sonuçları.	89
Şekil 4.7.	Farklı veri kümeleri için Otokodlayıcı, U-Net, RPCA, mRMR ve RF yöntemleriyle elde edilen özniteliklerin veri kümeleri arası tutarlılığı.	90
Şekil 4.8.	İBH ve KK veri kümelerinde Otokodlayıcı, U-Net ve RPCA yöntemleri ile elde edilen 10 biyobelirteç ve üç veri kümesi için ortak biyobelirteçler.	92
Şekil 4.9.	İBH ve KK veri kümelerinde mRMR ve RF yöntemleri ile elde edilen 10 biyobelirteç ve üç veri kümesi için ortak biyobelirteçler.	93

GİRİŞ

İnsan vücudunda yaşayan trilyonlarca mikrobiyal canlı, konağın sağlık ve hastalık durumlarının tanımlanmasında önemli bir rol oynar. Genellikle, bu canlıların sayısı insan hücrelerinden fazladır ve insan genomundan yaklaşık 150 kat daha fazla gen içerir [1]. Temel olarak Firmicutes, Bacteroidetes, Actinobacteria ve Proteobacteria gibi bakteri taksonlarından ve arkeler, mantarlar ve virüsler gibi diğer mikroorganizmalardan oluşan bağırsak mikrobiyomu, insan sağlığı üzerindeki derin biyolojik etkisi nedeniyle kapsamlı araştırmaların konusu olmuştur [2]. Mikrobiyom analizleri, günümüze kadar ağırlıklı olarak gözlemsel veya tanımlayıcı metodolojilere odaklanmıştır. Bu metotlar, belirli bir topluluk içindeki mikroorganizmaları tanımlamayı ve deneysel bulgular ile spesifik bir hastalıktan etkilenen belirli bir birey grubunda, belirli bir suşun varlığı veya yokluğu gibi çeşitli fenotipik özellikler arasında bağlantılar kurmayı içerir [3]. İnsan bağırsağı, konak hücrelerin ve yabancı organizmaların bir arada bulunması, iş birliği yapması ve rekabet etmesiyle karakterize edilen çok yönlü bir ekosistemdir. Bu ekosistem içerisinde, bu ortamda yaşayan organizmalar tarafından değiştirilen ve dönüştürülen bir araç olarak işlev gören çok sayıda besin metaboliti bulunur. Yeni nesil dizileme (NGS) ve metabolomik profillemadaki son gelişmeler, birçok farklı mikrop ve metabolit türünü tanımlamakta ve bunların göreceli bolluklarının ölçülmesinde etkili olmuştur [4].

İnsan ve bakteri genomunda yapılan çalışmalardaki gelişmeler ile bazı genlerin, maddelerin ve özelliklerin çeşitli hastalıkların teşhis veya tedavisinde kullanılabileceği tespit edilmiştir. Bu ayırt edici kilit özellikler biyolojik belirteç (biyobelirteç) olarak adlandırılır. Biyobelirteçler, organlarda ve hücrelerde fonksiyonel ve yapısal varyasyonların belirtileri olarak görülebilen, normal ve hastalıklı durumları ayırt edebilen, tedaviye verilecek yanıtları tahmin edip değerlendirebilen ve belirli hastalıkları ölçebilen göstergelerdir [5]. Ek olarak biyobelirteçler, ilaçların normal biyolojik süreçlere tepkilerini, hastalığın ilerleyişini ve tedavi yöntemlerini objektif olarak değerlendirebilir.

Bazı biyobelirteçler, sağlık koşullarındaki erken değişiklikleri tespit edebilen hastalık tanımlama belirteçleri olarak da hizmet eder [6, 7]. Pahalı olmayan, yüksek verimli dizileme teknolojilerinin sürekli gelişimi, mikrobiyal biyobelirteç keşif çalışmalarının yaygın kullanımına yol açmıştır. Biyobelirteç tespiti, metagenomik verilerden klinik uygulamaya geçmenin önemli bir yolu olarak düşünülebilir. Potansiyel biyobelirteçleri tespit etmek, hastalık gelişimini anlamak ve antibiyotik ve/veya probiyotik tedavileri tasarlamak için çok önemlidir. Mikrobiyal biyobelirteç keşfi, göreceli bollukları farklı fenotipler arasında farklılık gösteren spesifik operasyonel taksonomik birimleri (OTU) tanımlamayı amaçlamaktadır. Normal, sağlıklı bir mikrobiyota, yüksek bireysel çeşitlilik ve stabiliteye sahiptir. Bununla birlikte, mikrobiyota çeşitliliği kişiden kişiye oldukça değişiklik gösterir. Birçok hastalık, bu çeşitliliğin ve stabilitenin azalması olarak tanımlanabilen “disbiyozis” ile ilişkilidir. Mikrobiyota bileşimi, dışkı kıvamı ile yakından ilişkili olup antibiyotikler ve gıda gibi dış faktörlerden de büyük ölçüde etkilenir. İnvaziv olmayan hastalık teşhisi ve izlenmesi, klinisyenler için en çok arzulanan hedeflerden biridir [8]. Bu sebeple, hastalar için teşhis ve takibi kolaylaştıran invaziv olmayan yöntemlerin geliştirilmesine ihtiyaç duyulmaktadır.

Büyük miktarlarda biyolojik veriyi yorumlamak için üst düzey analiz tekniklerinin ve araçlarının kullanılması gerekir. Bu nedenle bu tür verilerde, büyük ve hacimli verilerin hızlı bir şekilde işlenmesi için tasarlanmış en gelişmiş hesaplama yöntemleri ve matematiksel ve istatistiksel teknikleri kullanmak önemlidir. NGS ile benzer şekilde metagenomik veri analizinin merkezinde yer alan istatistik, makine öğrenimi (ML) ve veri madenciliği alanları, son yirmi beş yılda önemli ölçüde ilerleme kaydetmiştir [6]. ML'nin hacimli verilerden yeni biyolojik hipotezler veya keşifler elde etmeye potansiyel olarak yardımcı olabileceği ve ilgili metagenom verilerindeki kalıpları ayırt etme ve tanımlama kapasitesine sahip olduğu düşünülmektedir. Bununla birlikte, her algoritmanın verimliliği, ham verilerden özellikler çıkarmak için alan bilgisinden yararlanmayı içeren özellik mühendisliğine bağlıdır. Manuel olarak yürütüldüğünde, ML için özellik mühendisliği süreci yanlışlıklara karşı hassas olabilir. Dahası, manuel özellik mühendisliği belirli problemlere uyarlanmıştır, bu da algoritmanın diğer problemleri ele almak için kullanılamayacağı anlamına gelir [9]. Derin öğrenme (DL) algoritmaları, farklı görevler için bilinen tekniğin en yüksek performanslı sonuçlarını

üretir. Bu algoritmalar, üstün performans gösterebilmek amacıyla, eğitim için büyük veri kümeleri gerektirir. Ayrıca, DL algoritmaları genellemeden yoksundur ve özellikle mikrobiyom verileriyle çalışırken, küçük veri kümeleri üzerinde eğitildiğinde aşırı öğrenmeden mustarıptır. Dahası, verilerin açıklamaları uzmanlar tarafından manuel olarak yapıldığında oldukça zaman alıcı olurken, kapsamlı, genelleştirilebilir bir model arayışı hesaplama maliyetlerini artırabilmektedir [10].

Metagenomik veri kümelerindeki en ayırt edici özelliklerin belirlenmesi, çeşitli nedenlerden dolayı zorlu bir görevdir. Bu nedenlerden biri de potansiyel biyobelirteçleri temsil eden öznitelik sayısının fazla olması olabilir, bu sorun genellikle "boyutsallığın laneti" olarak adlandırılır [1]. Biyobelirteç keşfi birçok çalışmada bir öznitelik seçimi (feature selection, FS) problemi olarak karşımıza çıkmaktadır [6, 11, 12]. Bir veri kümesindeki en önemli öznitelikleri tanımlamayı ve ilgisiz veya gereksiz olanları atmayı amaçlayan FS, biyolojik olarak ilgili özniteliklerin tanımlanması için önemlidir. Ayrıca, bu alanda yapılan çalışmalardan yalnızca birkaçı yüksek boyutlu ve az örnek bulunduran verilerle çalıştığı için moleküler verilerin temsil ettiği zorluklara odaklanabilmiştir. Nitekim, belirli bir örnek için ölçebildiğimiz özniteliklerin sayısı (mevcut teknolojilerle kolayca on milyonlara ulaşabilir) ile toplayabileceğimiz örnek sayısı (genellikle binler düzeyinde) arasında gittikçe açılan bir uçurum vardır. Bu yüksek boyutlu, düşük örneklem büyüklüğündeki veri kümeleri, istatistik ve ML yaklaşımlarının gücünü büyük ölçüde sınırlar [6]. FS prosedürlerinin gürbüzlüğü (veya sağlamlığı), yani verilerin küçük bozulmalarında aynı özellikleri muhafaza etme kabiliyetleri, yüksek boyutlu, düşük örneklemli ortamlar için oldukça önemli bir kıstastır. Gerçekte mevcut yaklaşımlar oluşturdukları modellerin tahmin hatasına odaklanma eğilimindedir ve ilgili öznitelikleri bulmak, optimal öngörü sağlayanları bulmaktan çok daha zordur. Mevcut algoritmalar tipik olarak oldukça istikrarsızdır ve genellikle aynı problemle ilgili farklı örnek kümeleri için oldukça farklı sonuçlar verir [6]. Bu yüksek değişkenlik, bu algoritmaların gerçekten alakalı öznitelikler yerine kendine özgü öznitelikleri yakaladığını ima eder. Bu durum, seçilen öznitelikler üzerine inşa edilen tahmine dayalı algoritmaların güvenilirliği konusunda şüphe uyandırır ve yeni biyolojik iç görüler elde etmek için bu özelliklerin yorumlanmasını engeller. FS sonucunun gürbüzlüğü, özellikle seçilen öznitelik alt kümelerinin sonraki analizleri veya doğrulamaları maliyetliyse arzu edilen

bir özelliktir. Bir alan uzmanının, FS algoritması tarafından seçilen biyobelirteçleri gerçek zamanlı polimeraz zincir reaksiyonu (RT-PCR) gibi tekniklerle doğrulaması yaygın bir uygulamadır. Ancak, farklı FS teknikleri, özniteliklerin farklı sıralanmasına neden olabilir, hatta aynı FS tekniği, yöntem parametrelerinin seçilen değerlerine bağlı olarak büyük ölçüde farklı sonuçlar üretebilir. Mevcut veri kümelerinin, genellikle mevcut eğitim örneklerinin sayısını aşan bir dizi öznitelikle tanımlanması sorunları daha da kötüleştirmektedir. Bu görece geniş veriler, özniteliklerin nihai sıralamasında daha da fazla çeşitliliğe yol açabilir. Bu sebeplerle metagenomik veriler özelinde, kararlı bir öznitelik seçme tekniği arzu edilir. Verileri oluşturan biyolojik sürecin, en azından karıştırıcı faktörler olmaksızın tüm hastalar için büyük ölçüde ortak olduğu varsayıldığından, belirli bir patoloji için hastaların farklı alt numunelerinde ilgili genlerin seçimi, hemen hemen aynı sonuçları üretmelidir [7].

Biyobelirteçlerin tanımlanması ve çeşitli bağırsak hastalıkları [13, 14], kanser [15], obezite [16] ve diyabet [17] gibi bağırsak mikrobiyomundaki değişimler ile ilişkili olduğu bilinen hastalıkların sınıflandırması birbiriyle yakından ilişkili iki problemidir. Ayrıca, hastalık sınıflandırılması veya tahmini, veri kümeleri üzerinde bir model oluşturma prosedürünü (seçim adımıyla ortaya çıkan önemli öznitelikleri girdi olarak kullanarak) ve ardından, oluşturulan model yoluyla numuneleri uygun hastalık alt tiplerine etiketleme prosedürünü ifade eder. Metagenomik verilerden elde edilen sonuçların tekrarlanabilirliği, klinik uygulamalar ve yanlış biyolojik sonuçları önlemek için önemlidir. Örnek büyüklüğündeki değişkenlik ve deneylere katılan denekler çeşitliliği tetikler ve bu da biyobelirteç algılama algoritmalarının sonucunu büyük ölçüde değiştirebilir. Bu zorluklar mevcut biyobelirteç tanımlama yöntemlerinin gerçek biyobelirteçleri seçmesini engelleyen, ciddi bir tutarsızlık sorununa işaret etmektedir. Örneğin Chuang vd. [18], iki gen ekspresyon çalışması [19, 20] tarafından meme kanseri için potansiyel biyobelirteçler olarak önerilen 70 genden yalnızca üçünün ortak olduğunu bildirmiştir. Bu nedenle numunelerde bulunan doğal çeşitlilikten bağımsız olarak, biyolojik verilerden elde edilen sonuçların tekrarlanabilirliğini sağlayan güçlü bir biyobelirteç algılama algoritması geliştirmek, doğru biyolojik ifadelerin çıkarılmasında ve bu sonuçların başarılı klinik çalışmalarda kullanılmasında kritik bir rol oynamaktadır [12].

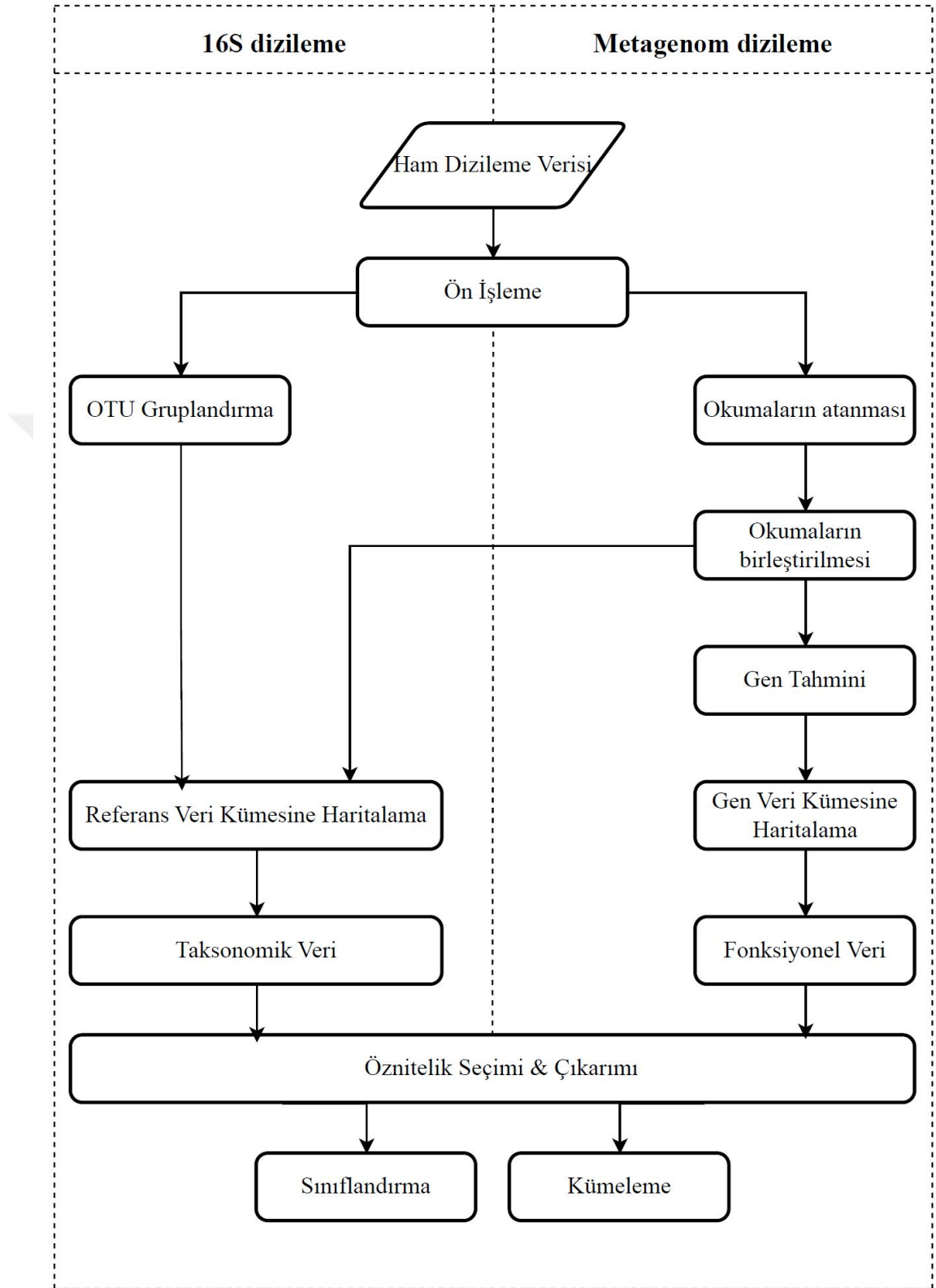
Bu motivasyonla, metagenom verilerindeki yüksek boyutluluk, veri standardizasyonu ve sınırlı verilerle ilgili zorluklara çözüm bulmak amacıyla yeni algoritmalar geliştirilmiştir. Otokodlayıcı, U-Net ve Üretici Çekişmeli Ağ (GAN) tabanlı DL yaklaşımlarıyla oluşturulan modeller, Aterosklerotik kardiyovasküler hastalık (ACVD), Ankilozan spondilit (AS), Kolorektal kanser (KK), İltihaplı bağırsak hastalığı (İBH), Crohn hastalığı (CH), Graves hastalığı (GH), Karaciğer Sirozu (KS), Multipl skleroz (MS), Non-Alkolik Yağlı Karaciğer Hastalığı (NAYKH), Romatoid Artrit (RA), Tip 2 diyabet (T2D) ve Ülseratif kolit (ÜK) için bağırsak mikrobiyom verilerini içeren veri kümeleri üzerinde doğru ve sağlam sınıflandırma performansı ölçümü için kullanılmış, ayrıca KK ve İBH için kararlı biyobelirteçlerin keşfi için değerlendirilmiştir.

Tez çalışması 5 bölümden oluşmaktadır. 1. Bölüm, mikrobiyomun sağlık ve hastalık ile ilişkisi ve mikrobiyom verilerinin analizi için kullanılan yöntemleri özetleyerek literatür hakkında genel bilgileri, bu alanda karşılaşılan problemleri ve tezin amaç ve önemini aktarır. 2. Bölüm, tez çalışmasında kullanılan DL mimarileriyle ilgili detaylı bilgiler ve literatürde bu mimarileri kullanan mikrobiyom çalışmalarına ayrılmıştır. 3. Bölüm, tez çalışmasında mikrobiyom verilerinde sınıflandırma performansını iyileştirmek için geliştirilen Otokodlayıcı, U-Net ve GAN tabanlı yöntemlere ait detaylı bilgiler ve analiz sonuçlarını içermektedir. 4. Bölüm, bağırsak mikrobiyomundan çeşitli hastalıklara ilişkin kararlı biyobelirteçlerin keşfi için geliştirilen yaklaşımların detaylarını içermektedir. Son olarak, 5. Bölüm tez çalışmasında geliştirilen tüm metotlar ve elde edilen sonuçlar hakkında tartışma ve gelecekteki çalışmalar için önerileri aktarır.

1. BÖLÜM

GENEL BİLGİLER VE LİTERATÜR ÇALIŞMASI

Gastrointestinal (GI) sistemde bulunan karmaşık ve dinamik bir mikroorganizma topluluğu olan insan bağırsak mikrobiyomu, insan sağlığı ve hastalıklarında önemli bir bileşen olarak ortaya çıkmıştır [21–23]. Trilyonlarca bakteri, virüs, mantar ve diğer mikroorganizmalardan oluşan bağırsak mikrobiyomu, sindirim, bağışıklık sistemi ve hatta sinirsel fonksiyon dahil olmak üzere çok çeşitli fizyolojik süreçleri etkiler [24, 25]. Geçmişte bakteri topluluklarının araştırılmasının sınırlı kalmasının sebebi, mikropların %90'ından fazlasının tanımlanamamış olmasıdır [1]. Günümüzde ise dizileme teknolojilerindeki son gelişmeler bu sınırlamaların üstesinden gelmiş ve araştırmacıların mikrobiyal kolonilerin taksonomik bileşimi ve fonksiyonel kapasitesi hakkında fikir sahibi olmasını sağlamıştır [1, 2]. Bir mikrobiyoma ait bakteriyel çeşitliliğin ortaya konması için genellikle görece düşük maliyetli olan, komünitedeki tüm 16S ribozomal RNA (rRNA) genlerinin topluca dizilendiği “16S dizilemesi” olarak kısaca isimlendirilebilecek bir yaklaşım kullanılmaktadır. Komünitedeki mikrobiyal çeşitliliğin ötesinde tüm genetik yapı (tüm bakteriyel genler, DNA virüsleri, ökaryotlar, dolayısıyla buradan çıkarılabilecek protein, yolak ve metabolizma bilgileri) ortaya konmak istediğinde kısaca “metagenom dizileme” olarak adlandırılabilir, ortamdaki tüm DNA'nın izole edilip topluca yüksek çıktılı olarak dizilendiği görece olarak daha yüksek maliyetli yaklaşıma baş vurulmaktadır. Şekil 1.1' de dizileme yöntemlerine ait süreçleri içeren bir akış şeması verilmiştir. 16S ve metagenom dizilemeleri analizi sonucunda, her bir mikrobiyom örneğine ait yüzler-binler seviyesinde taksonomik veya fonksiyonel grubun, varlık ve nicelik bilgisi elde edilebilmektedir. Son gelişmeler bağırsak mikrobiyomunun, hastalık tanısı ve prognozu için bir biyobelirteç olarak potansiyelini vurgulamıştır [26]. Disbiyoz olarak tanımlanan bağırsak mikrobiyomunun



Şekil 1.1. Dizileme yöntemlerinin süreç adımlarını içeren iş akış şeması.

bileşimindeki ve işlevindeki bozulmalar, İBH ve KK'den nörodejeneratif bozukluklara ve karmaşık kronik hastalıklara kadar çeşitli hastalıkların patogeneğinde rol oynamaktadır [27–31].

1.1. Mikrobiyomun Sağlık ve Hastalık ile İlişkisi

Vücudumuzdaki mikropların yaklaşık olarak 38 milyar hücre oluşturduğu ve bu sayının insan hücreleri sayısının yaklaşık 1.3 katı olduğu tahmin edilmektedir [32]. İnsan genomundan 100 kat daha fazla benzersiz gen kodladıkları öne sürülen mikropların çoğu bağırsakta bulunur. Bu mikroplar insan fizyolojisi ve beslenmesi üzerinde derin bir etkiye sahiptir ve insan yaşamı için çok önemlidir. Bağırsak mikroplarının insan sağlığı üzerindeki etkisini anlamak için mikrobiyal bağırsak topluluğunun içeriğini, çeşitliliğini ve işleyişini deşifre etmek gerekir. 16S dizisine dayalı yöntemler, Bacteroidetes ve Firmicutes'in bilinen filogenetik kategorilerin %90'ından fazlasını oluşturduğunu ve bağırsak mikrobiyotasına hâkim olduğunu ortaya çıkarmıştır. Çalışmalar ayrıca, bağırsak mikrobiyomunda, sağlıklı bireyler arasında önemli bir çeşitliliğe işaret etmiştir. Bu farklılık özellikle bebekler arasında belirgin olmasına rağmen, bağırsak mikrobiyomu yaşamın ilerleyen dönemlerinde daha benzer filumlara yakınsamaktadır [33].

İBH, GI sistemde kronik inflamasyonu içeren bir bozukluk olarak tanımlanır [34]. İBH'nin iki ana alt tipi vardır: ÜK, rektumdan yayılan ve kolon boyunca uzanan diffüz mukozal inflamasyon ile belirli bir desen sergilerken, CH sürekli olmayan atlamalı lezyonlarla tüm GI sisteminin herhangi bir yerinde ortaya çıkma potansiyeline sahiptir [34–36]. Ayrıca, yine bir bağırsak hastalığı olan KK, son dönemlerde yaygınlığı artmış olan en ölümcül kanser türlerinden biri olarak tanımlanabilir [27]. Birçok çalışma, İBH ve KK'nin ilişkili olduğunu öne sürmüştür, ayrıca İBH'li bireylerin sağlıklı insanlara göre KK geliştirme olasılığı iki ila altı kat daha fazladır [37]. Enflamasyon, hastalığın süresi, başlangıç yaşı, çocuklukta başlangıç, lezyonların yayılması ve sklerozan kolanjit gibi faktörler, bir bireyin kansere karşı duyarlılığını artıran ve dolayısıyla İBH'li bireylerde KK gelişme olasılığına katkıda bulunan faktörler olarak kabul edilir [38]. Son birkaç yılda, çok sayıda çalışma insan bağırsak mikrobiyomu ile çeşitli hastalıklar arasında önemli korelasyonlar olduğunu göstermiştir [39, 40]. Giderek artan bir kanıt grubu, mikrobiyomun inflamatuvar yolları etkileyerek veya kanserojen mikrobiyal metabolitler

üretmek KK'nin gelişimine veya ilerlemesine katkıda bulunabileceğini öne sürmektedir [41]. Benzer şekilde, İBH'li bireylerde mikrobiyal kompozisyonun hem miktarında hem de kalitesinde değişiklikler belgelenmiştir [42]. Ayrıca çok sayıda çalışma, GI bozukluklar arasındaki bağlantıyı ve bunların bir formdan diğerine geçiş potansiyelini vurgulamaktadır. Örneğin, her iki ana İBH türündeki kronik inflamasyon, KK, intestinal lenfoma ve ince bağırsak adenokarsinomu gibi diğer GI hastalıklarının gelişmesine yol açabilir [43]. Bu kanserlerin artan sıklığı, İBH'li bireylerin hayatta kalmalarını önemli ölçüde etkileyebilir, genel kalitelerini ve yaşam sürelerini azaltabilir. Bu faktörler nedeniyle, doğru ve zamanında tanı, büyük önem taşımaktadır.

İBH için güncel klinik tanı yöntemleri arasında kan testleri, kolonoskopi ve MRI bulunur. Ancak, belirgin olmayan semptomlar nedeniyle, İBH'nin yanlış teşhisi klinik uygulamada yaygın bir olgudur [44, 45]. Klinik araştırmalar, KK'nin prognozunu tanı zamanlamasına bağlı olduğunu ve daha erken tespitin hastalara daha önemli bir hayatta kalma avantajı sağladığını ortaya koymuştur [46]. İBH'ye benzer şekilde, Dışkıda Gizli Kan Testi (FOBT) ve kolonoskopi gibi geleneksel yaklaşımlar KK tespiti için belirli sınırlamalara sahiptir. Açıklamak gerekirse, FOBT yöntemi hem yanlış pozitif hem de yanlış negatif sonuçlarla mücadele eder [47]. Buna paralel olarak, kolonoskopi sadece pahalı olmakla kalmaz, aynı zamanda hastalar için rahatsızlık yaratabilir ve perforasyon ve kanama riskleri taşıyabilir [48]. Dahası, kolonoskopi yalnızca hastane ortamında gerçekleştirilebilen bir tıbbi muayenedir. Ancak, bağırsak mikrobiyomu analizi hastalara ev test kiti ile kendi evlerinde örnek verme şansı sağlayabilir. Bu nedenle, gerekli örnekleri toplamak nispeten daha kolaydır ve böylece hastane masrafları azaltılabilir. Yukarıda belirtilen sınırlamalar göz önüne alındığında, dışkı örneklerinden bağırsak mikroplarının analiz edilmesinin, GI bozuklukları tespit etmek için daha etkili ve invaziv olmayan bir yaklaşım sunabileceği fikri destek kazanmaktadır [49].

GI hastalıklar haricinde de bağırsak mikrobiyomu ile ilişkilendirme çalışmaları yapılan pek çok hastalık bulunmaktadır. Örneğin, mikropların insan fizyolojisi ve beslenmesi üzerinde derin bir etkiye sahip olması, ayrıca bağırsaktaki mikropların, yiyeceklerden enerji toplanmasına katkıda bulunması sebepleriyle bağırsak mikrobiyomu değişikliklerinin obezite ve diyabet ile ilişkili olabileceği düşünülmüştür [1]. Ayrıca kronik karın ağrısı ve dışkı düzeninde değişiklik gibi belirtilerle ortaya çıkan, patogenezi

Tablo 1.1. Farklı hastalıklar için bağırsak mikrobiyomunda gözlemlenen değişiklikler.

Hastalık	Mikrobiyom Değişikliği	Referans
NAFLD	<i>Firmicutes</i> ↑, <i>Bacteroidetes</i> ↑	[50]
Siroz	<i>Firmicutes</i> ↑, <i>Bacteroidetes</i> ↑, <i>Streptococcus</i> spp. ↓, <i>Veillonella</i> ↓	[50]
T1D	<i>Bacteroidetes</i> ↑, <i>Clostridia</i> ↑, Bakteriyel çeşitlilik ↓	[51]
T2D	<i>Bacteroidetes</i> ↑	[51]
Otizm	<i>Clostridial</i> türler ↑, <i>Sutteralla</i> ve <i>Desulfovibrio</i> türleri ↑	[51]
HBS	<i>Dialister</i> spp. ↑, <i>Faecalibacterium prausnitzii</i> ↑, <i>Streptococcus</i> ↓, <i>Bacteriodes fragilis</i> ↓, <i>Clostridium difficile</i> ↓	[52]
ÜK	<i>Bacteriodes fragilis</i> ↑, <i>Dialister</i> spp. ↑, <i>Roseburia</i> spp. ↑, <i>Clostridium difficile</i> ↓	[52]
CH	<i>Dialister</i> spp. ↑, <i>Clostridium difficile</i> ↑, <i>Streptococcus</i> spp. ↑, <i>Ruminococcus gnavus</i> ↑, <i>Bacteriodes fragilis</i> ↓, <i>Faecalibacterium prausnitzii</i> ↓	[52]
İBH	<i>Enterobacteriaceae</i> ↑, <i>Rikenellaceae</i> ↓, <i>Streptococcus</i> spp. ↓, <i>Faecalibacterium prausnitzii</i> ↓, <i>Akkermansia muciniphila</i> ↓	[52]
KK	<i>Fusobacterium</i> ↑, <i>Porphyromonas</i> ↑, <i>Clostridia</i> ↓	[52]
Obezite	<i>Firmicutes</i> ↑, <i>Bacteroidetes</i> ↓	[53]
RA	<i>Bifidobacteria</i> ↓, <i>Porphyromonas</i> ↓	[53]

tam olarak anlaşılamamış, Huzursuz bağırsak sendromu (HBS) [54] ve çölyak hastalığı [55], karmaşık bir genetik otoimmün bozukluk olarak tanımlanan kronik ve tedavi edilemeyen bir hastalık olan RA [29], erken belirtileri arasında görme sorunları, yürümede zorluk, konuşma problemleri bulunan ve merkezi sinir sistemini etkileyen kronik bir hastalık olan MS [31] ve nörodejeneratif bir hastalık olan, hafızayı, düşünmeyi ve davranışı etkileyen Alzheimer hastalığı [30] da dahil olmak üzere hastalıkların pek çoğunu bağırsak mikrobiyomuyla ilişkilendiren çalışmalar mevcuttur. Tablo 1.1’de literatürdeki çalışmalarda farklı hastalıklar için bağırsak mikrobiyomunda gözlemlenen değişiklikler listelenmiştir.

1.2. Mikrobiyom Veri Analizinde Kullanılan Yöntemler

Yüksek verimli dizileme teknolojilerindeki gelişmeler, dünyamızda yaşayan mikrobiyal toplulukların karakterizasyonu için ortam hazırlamaktadır. Metagenomik uygulama sayısındaki artış, maliyetlerdeki düşüş ve mikrobiyota işlevlerinin derinlemesine anlaşılması isteği, genetik veri üretiminde hızla artışa yol açmaktadır. 2025 yılında dizileme yoluyla üretilen toplam veri miktarının astronomi, YouTube veya Twitter ile aynı veya daha fazla olması tahmin edilmektedir [56]. Metagenomik, bu verilerin önemli bir alt kümesine katkıda bulunacaktır. Tek bir mikrobiyom çalışması, yüzlerce gigabyte veya daha fazla ham dizileme verisi içerebilir hatta işleme sırasında, farklı formatlarda ara sonuçlar üretilmesi gerektiğinden, bu değer birçok kez katlanabilir. Bu nedenle, bu miktarda veriyi işleyebilen algoritmalara ve mimarilere sahip olmak önemlidir [57]. Geleneksel yöntemler, metagenomik verilerin benzersiz doğası nedeniyle engellerle karşılaşmaktadır. İşte tam bu noktada, yapay zekanın günümüzde iyi bilinen bir dalı olan Derin Öğrenme (Deep Learning, DL) devreye girmektedir. DL tabanlı teknikler, mevcut yöntemleri tamamlayıp ve mikrobiyom araştırmalarında yeni yollar açmaktadır. Bağırsak mikrobiyomu, hastalık tanısı ve prognozu için hayati bilgiler sağlayarak insan sağlığında önemli bir rol oynar. Ancak, referans katalogları, seyreklik ve çeşitlilik gibi faktörler nedeniyle metagenomik verilerin analizi zor olmaya devam etmektedir. DL, son teknoloji mikrobiyom hatlarını tamamlayan yeni ve umut verici yaklaşımları mümkün kılar. DL tabanlı yöntemler, yeni patojen tespiti, bağışıklık ölçümü, hasta sınıflandırması ve hastalık tahmini dahil olmak üzere mikrobiyom analizinin hemen hemen tüm yönlerini ele alabilir. Bu yöntemlerin önemli bir yönü de tahmini modeller üretmenin ötesinde, yorumlanabilirlikleridir [58]. Mikrobiyom araştırmalarında büyük miktarda veri, bol miktarda öznetelik üretilmesi nedeniyle, Makine öğrenmesi (Machine Learning, ML) yöntemleri bu alanda yaygın olarak kullanılmaktadır. Hasta fenotipini tahmin etme görevi, çeşitli ML modelleri kullanılarak ele alınabilir. Bu algoritmalar, artan sayıda veri kümesiyle, çeşitli özelliklerine göre örnekleri sınıflandırmak için verilerden önemli desenler öğrenebilir ve çıkarabilir.

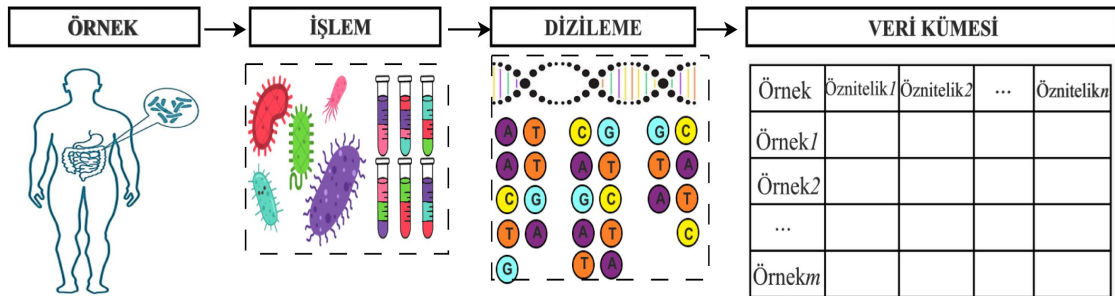
1.2.1. Mikrobiyom ve Makine Öğrenmesi

Yapay zekanın bir dalı olan ML, farklı alanlardan karmaşık veri örüntülerini tanımlamak

için oldukça etkili bir çözüm olarak görülmektedir. ML için temel terminoloji ve gerekliliklere değinmek gerekirse: (1) veri kümesi (eğitim ve test); (2) örnek; (3) öznitelik; (4) sınıf etiketi ve (5) çapraz doğrulama olarak özetlenebilir. Belirtilen terimlerin bilgisi, geliştirilen ML modellerinin ardında yatan bağlamların anlaşılmasını sağlar. Her birinin kısa açıklaması şu şekildedir: (1) Veri kümesi, özniteliklere karşı örnekler matrisidir. Eğitim veri kümesi, bir sınıflandırıcı geliştirmek ve öğrenme modelini eğitmek için kullanılan bir veri alt kümesidir. Test veri kümesi, eğitimin bir parçası olmayan ve sınıflandırıcı performansını analiz etmek amacıyla sınıflandırıcının hatasını tahmin etmeye yardımcı olan veri alt kümesidir. (2) Örnek (vaka, kayıt veya numune olarak da bilinir), modeli eğitmek veya test etmek için kullanılır. Örneğin KK ile ilgili çalışmalarda, örnekler tümör veya hasta/sağlıklı kontrol numunelerinin sayısıdır. (3) Öznitelik (nitelik, değişken veya özellik olarak da bilinir), bir örneği tanımlayan sayısal veya kategorik bir niceliktir. Bir durumu temsil eden bir dizi öznitelik vektörü, öznitelik uzayı olarak adlandırılır. Kanser biyolojisi ile ilgili örnek verecek olursak öznitelikler; görüntüleme verileri, genomik veriler (tek nükleotid polimorfizmi), transkriptomik veriler (mRNA ve miRNA ifadesi), proteomikler, metabolomik veriler, hastanın fenotipini tanımlayan klinik parametreler veya dışkıdan elde edilmiş bağırsak mikrobiyomu numuneleri olabilir. (4) Sınıf etiketi, bir örneğin sınıf kategorisinin tanımıdır (Hasta, sağlıklı kontrol veya hastalığın farklı evrelerinden/çeşitlerinden farklı etiketler). (5) Çapraz doğrulama ise (Cross Validation, CV), bir sınıflandırıcının hatalarını ve genelleme yeteneklerini tahmin eder. K-fold, birini dışarıda bırakma (Leave one out, LOO), holdout, bootstrapping vb. gibi çeşitli CV yöntemleri mevcuttur. K-fold CV' de, veri kümesi k ayrı alt kümeye bölünür, bu aşamadan sonra sınıflandırıcı, her alt küme test veri kümesi olarak kullanılıncaya kadar $k - 1$ alt küme üzerinde eğitilip 1 alt kümede test edilir. k , veri kümesinin örnek sayısına eşitse, LOO-CV prosedürü olarak adlandırılır. Alt kümelerinin ortalama doğruluğu, sınıflandırıcının doğruluk tahmini olarak kabul edilir [59].

Mikrobiyom araştırmalarında ML, fenotipleme (yani, çevresel veya konak fenotipini tahmin etme), mikrobiyal özellik sınıflandırması (yani, mikrobiyotanın bolluğunu, çeşitliliğini veya dağılımını belirleme), mikrobiyomun bileşenleri arasındaki karmaşık fiziksel ve kimyasal etkileşimleri inceleme ve mikrobiyomun bileşimindeki değişiklikleri

izleme gibi görevleri ele almak için uygulanır [60]. ML algoritmaları, literatürde genellikle sınıf etiketlerinin eğitim örnekleriyle ilişkilendirilmesine göre denetimli öğrenme, denetimsiz öğrenme ve pekiştirmeli öğrenme olarak sınıflandırılır. Denetimli ML algoritmaları, örneklerin sınıf etiketinin hali hazırda mevcut bulunduğu sınıflandırma veya regresyon problemleri için kullanılır. Örneğlerin sınıf etiketi bilgisinin mevcut olması durumunda, kümeleme, yoğunluk tahmini veya yüksek boyutluluğun azaltılması gibi görevler için denetimsiz ML algoritmaları uygulanır. Pekiştirmeli öğrenmenin diğer öğrenme yöntemlerinden en büyük farkı, bulunduğu çevreyle girdiği etkileşim sonucu süreç boyunca maksimum ödülü kazanmak için optimal aksiyonları belirlemesidir. Özetle, pekiştirmeli öğrenmenin davranışçılıktan esinlenen, öznelerin bir ortamda en yüksek ödül miktarına ulaşabilmesi için hangi eylemleri yapması gerektiğiyle ilgilenen bir ML yaklaşımı olduğu söylenebilir. ML algoritması tarafından oluşturulan modeller / sınıflandırıcılar, yeni örnekleri veya hastaların klinik fenotipini tahmin etmek için bağımsız yürütülebilir sistemler olmalıdır. Mikrobiyom verilerinden hastalık teşhisi için model geliştirmenin ilk adımı olan veri üretimine yönelik basitleştirilmiş bir iş akışı Şekil 1.2’de gösterilmektedir. Her şeyden önce, çeşitli ortamlar, insanlar veya konaklarla ilişkili bakteri topluluklarından dizileme verilerinin toplanması gerekir. Veri kümesi genellikle sınıf etiketleri (klinik fenotip) ve m satır (hastalar / örnek) ve n sütundan (özellikler / öznitelikler) oluşan bir matristir. Mikrobiyom verilerinden gelen öznitelikler söz konusu olduğunda, n çoğunlukla m’den büyüktür.



Şekil 1.2. Mikrobiyom verilerinden hastalık teşhisi için model geliştirmenin ilk adımı olan veri üretimine yönelik basitleştirilmiş bir iş akışı.

1.2.1.1. Klasik Makine Öğrenmesi Yöntemleri

Klasik ML yöntemleri arasında en popüler yöntemlerden olan lojistik regresyon (LR),

rastgele orman (RF) ve destek vektör makineleri (SVM) yöntemlerinin mikrobiyom verilerinde iyi performans gösterdiği bilinmektedir ancak son zamanlardaki çalışmalarda, yeni yöntemlerle kıyaslama aracı olarak evrilip kullanımları azalmıştır [60]. SVM'ler [61], giriş verilerini iki ayrı sınıfa ayırabilmek için, vektör uzayında iki sınıfın veri noktaları arasında, mümkün olan en büyük marja sahip bir hiper düzlem hesaplayarak çalışır. Hiper düzlem, içinde bulunduğu vektör uzayından bir boyutu daha az olan bir alt uzay olarak tanımlanır. Böyle bir hiper düzlem varsa, veri doğrusal olarak ayrılabilir. Temsili doğrusal olarak ayrılabilir veriler üzerinde bir SVM eğitmek, %100 doğrulukla mükemmel bir sınıflandırma ile sonuçlanacaktır. Ne yazık ki, gerçek dünya verilerinin çoğu, doğrusal bir şekilde ifade edilemeyen boyutlar, aykırı değerler veya karmaşık bağımlılıklar nedeniyle tamamen doğrusal olarak ayrılabilir değildir. Bu nedenle, çoğu SVM uygulaması, bu tür hatalara izin verir ve aykırı değerlerin getirdiği hatayı en aza indirmeye çalışırken aynı zamanda doğrusal olarak ayrılabilir veri noktaları arasındaki marjı en üst düzeye çıkarmaya çalışır.

Karar ağaçları, her düğümün bir öngörüye göre veri kümesini iki alt kümeye bölündüğü bir ağaç yapısıdır. Bu işlem, veri kümesini daha küçük alt kümelere bölen alt düğümlerle tekrarlanabilir. Bir karar ağacı, ağacın yapraklarına sınıf etiketleri atayarak bir sınıflandırıcı olarak kullanılabilir. Düğümler, ağacın her seviyesindeki tahmin değerini en üst düzeye çıkaracak şekilde oluşturulmalıdır, yani boyutlar ve değerler veri noktalarını hedef sınıflara en iyi şekilde ayıracak şekilde seçilmelidir. Sınıflandırıcı olarak bir karar ağacı kullanmak, ağaç verimli bir şekilde oluşturulabildiğinden sınıflandırıcı oluşturmanın en hızlı yollarından biridir, ancak karar ağaçları birçok gerçek dünya problemi için çok basittir ve aşırı uyuma (overfitting) eğilimlidir, bu da ağacın eğitim verilerini sınıflandırabilmesine rağmen yeni verileri doğru bir şekilde sınıflandıramayacağı anlamına gelir [57].

RF yöntemi [62], karar ağacı sınıflandırıcısının eksikliklerini gidermeye çalışır. Bir değil, birkaç (genellikle yüzlerce veya binlerce) karar ağacı kullanır ve ardından nihai sınıflandırma kararını çoğunluk oyuna dayandırır. Aynı boyutları kullanarak, aynı veri kümesinden birden fazla karar ağacı oluşturmak, sınıflandırma performansını iyileştirmeden aynı veya çok benzer ağaçlarla sonuçlanacağından, ağaçlar için özel bir yapılandırma mantığı kullanılmalıdır. Bu yapılandırma mantığı, her ağaç için toplam

verinin yalnızca bir alt kümesini kullanmaktan ve tahminlerin seçilebileceği boyutları, oluşturulan ağaçların her seviyesindeki toplam boyutların rastgele bir alt kümesiyle sınırlamaktan oluşur. Ağaçların bu şekilde oluşturulmasının, çok sayıda ağaç kullanılsa bile aşırı uyum riskini önlediği söylenebilir.

Gradyan artırma (Gradient boosting, GB), varyansı azaltmak için kullanılan yaklaşımlardan biridir. RF'nin aksine, GB ağaçlar modelinin ağaçları sırayla oluşturulur ve her ağaç, önceki ağacın hatasını azaltmayı amaçlar. Model/tahmin, sabit bir değerle başlatılır ve her bir yinelemede bir ağaç eğitilir. Her yineleme sonunda, yanlış sınıflandırılan örneklerin ağırlıkları artırılır ve doğru sınıflandırılan örneklerin ağırlıkları azaltılır. Dolayısıyla, her bir ardışık sınıflandırıcı, önceki adımlarda sınıflandırılması zor olan örnekler odaklanır. Yinelemeler sonunda sınıflandırıcı tahminleri, ağırlıklı çoğunluk oyu ile nihai bir tahminde birleştirilir [63]. GB yöntemlerinin çeşitli hastalıklar üzerinde karar ağaçlarına göre daha iyi performans gösterdiği ve RF ile yakın sonuçlar verdiği gösterilmiştir [64].

LR [62] giriş vektörünün ağırlıklı bir toplamını hesaplar ve daha sonra bu toplama "lojistik fonksiyon" olarak da bilinen sigmoid fonksiyonunu uygular. Sigmoid fonksiyonunun sonucu, bir veri noktasının belirli bir sınıfa ait olma olasılığı olarak yorumlanabilir. Bir LR sınıflandırıcısının eğitimi, giriş vektörü ağırlıklarını, çıktı fonksiyonunun eğitim veri kümesiyle en iyi şekilde eşleşmesi için ayarlamakla gerçekleşir. Sınıflandırıcıdaki lojistik fonksiyonuna tek giriş, girdi değerlerinin doğrusal bir birleşimi (ağırlıklandırılmış toplam) olduğundan, sınıflandırıcının iyi performans göstermesi için, giriş verilerinin çoğunlukla doğrusal olarak ayrılabilir olması gerekir. LR sınıflandırıcısı ile gerçekleştirilen hesaplamalar, aktivasyon fonksiyonu olarak sigmoid fonksiyonunu kullanan bir sinir ağının tek bir nöronundaki hesaplamalarla aynıdır [57]. LR, basitliği nedeniyle mikrobiyomu inceleyen çalışmalarda, özellikle FS ve kıyaslama için yaygın olarak kullanılan doğrusal bir sınıflandırma modelidir. LR sınıflandırıcıları doğrusal karar sınırlarına sahip olmalarına rağmen, mikrobiyomla ilişkili konak tahmininde doğrusal olmayan muadillerine benzer sonuçlar elde edebilir [65]. Lasso ve Elastik ağlar (ENet) gibi LR yöntemleri, fenotip gibi bir çıktıyı, girdilerin doğrusal bir kombinasyonu olarak modelleyerek bu yöntemlerin yorumlanmasını basit hale getirir. Bu yöntemler son zamanlarda konak disbiyoz tahmini çalışmalarında

kullanılmış olup, RF gibi diğer yöntemlerle karşılaştırılabilir sonuçlar vermiştir [66].

1.2.1.2. Boyut İndirgeme için Kullanılan Makine Öğrenmesi Yöntemleri

Mikrobiyom veri kümeleri, genellikle örnek sayısından daha fazla sayıda özneliğe sahiptir. Örnek sayısına göre çok sayıda özneliğe sahip veriler üzerinde eğitilen ML modelleri, karar sınırları oluşturmak için kullanılan olası ilişkilerin çoğunun sahte olması nedeniyle aşırı uyuma eğilimlidir. Ayrıca, ML modellerini eğitmek hesaplama açısından maliyetli olabilir, bu durumda eğitim verilerine aşırı uyum sağlama olasılığı daha düşük, daha basit modeller kullanılması ya da alternatif olarak boyut azaltma/indirgeme yöntemlerinin tercih edilmesi önerilir [63]. Boyut indirgeme yöntemleri, orijinal verileri daha az özneliğe sahip bir veri kümesine dönüştüren öznelik çıkarımı ve orijinal özneliklerin bir alt kümesini seçen FS yaklaşımlarını içerir [63]. FS, en bilgilendirici özneliklerin tanımlanmasına odaklanarak girdi değişkenlerinin sistematik olarak azaltılmasını içerir, böylece sınıflandırma modellerinin duyarlılığı artar ve yanıltıcı (ilgisiz ve gereksiz) öznelikler azaltıldığından daha hızlı ve daha doğru tahminler sağlar [67,68]. Dahası, ölçülecek girdi değişkenlerinin sayısını ve verilerdeki gürültüyü azaltarak aşırı uyum riskini en aza indirir. Son olarak, FS yöntemleri, potansiyel biyobelirteçler olarak hizmet eden metagenomik veri kümelerindeki özneliklerin bir alt kümesini tanımlamada çok önemlidir [69].

Denetimsiz düzenleme yöntemleri boyutluluğu azaltır ve verileri insan yorumu için basitleştirir. Bu algoritmalar görselleştirmeler veya sözde gösterimler oluşturmak için uygundur, mevcut özneliklerin doğrusal veya doğrusal olmayan bir kombinasyonunu hesaplayarak, giriş verilerinin sıkıştırılmış bir gösterimini oluşturur. Temel bileşen analizi (PCA) ve temel koordinat analizi (PCoA) gibi doğrusal yöntemler, mikrobiyal toplulukları görselleştirmek ve karşılaştırmak için popüler araçlardır. Bu yöntemler orijinal veri kümesindeki mevcut öznelikleri daha düşük boyutlu bir uzayda yeni bir öznelik kümesi ile temsil ederek bir ön işleme adımı görevi görür.

FS yöntemleri, öznelik çıkarım yöntemlerinin aksine orijinal veri kümesindeki öznelikleri korur ve farklı kurallarla yalnızca ilgili özneliklerin bir alt kümesini seçmeye odaklanır. Popüler FS yöntemlerinden biri olan Minimum Fazlalık- Maksimum

Alaka (mRMR) yöntemi mikrodizi gen ifadesi verileri için geliştirilmiş olup, sıklıkla mikrobiyom karakterizasyonu çalışmalarında kullanılmaktadır [63]. Bu yöntemde öznitelikler, alaka düzeyi, bilgi kazanımı veya F testi kullanılarak seçilir. İlk seçilen öznitelik, en yüksek alaka kriterine sahip olanıdır. Boyut indirgeme yöntemlerinin birçoğu mikrobiyom verilerinde biyobelirteç keşfi amacı ile de kullanılmaktadır.

1.2.1.3. Biyobelirteç Keşfi için Kullanılan Makine Öğrenmesi Yöntemleri

Günümüzde biyobelirteçler olarak adlandırılan belirli bir biyolojik durumun göstergesi olan maddelerin tanımlanması, tıp, biyomedikal ve biyoinformatik alanları için önemli bir araştırma konusudur. ML açısından bakıldığında biyobelirteç tespiti, amacın hasta ve kontrol grubu arasındaki farkı en iyi açıklayan bir dizi öznitelik (belirteç) bulmak olduğu bir FS problemi olarak ifade edilebilir [6, 12].

FS, alakasız veya gürültülü öznitelikleri ortadan kaldırarak daha güçlü sınıflandırıcılar elde etmek, orijinal özniteliklerin yalnızca küçük bir alt kümesini kullanarak daha kompakt ve daha hızlı modeller geliştirmek ve yalnızca alakalı bir alt kümeye odaklanarak yeni bilgilerin keşfi için kullanılabilir özniteliklerin tespiti gibi bir dizi avantaj sunar. Bu sayede gelecekteki klinik testler, potansiyel olarak daha az belirteçle ve daha düşük bir maliyetle oluşturulabilir [11]. FS teknikleri, sınıflandırıcıyla nasıl etkileşim kurduklarına bağlı olarak üç kategoriye ayrılır [70]. Filtreleme yöntemleri, doğrudan veri kümesi üzerinde çalışır ve çıktı olarak bir öznitelik ağırlıklandırma sıralamasını veya alt kümesini bulur. Sarmalayıcı yöntemler, modelin sonucuna göre (ör. Eğitim kümesinin CV sınıflandırma performansı) öznitelik alt kümeleri arasında bir arama gerçekleştirir. Sınıflandırıcı ile etkileşimde olduğundan genellikle filtreleme yöntemlerinden daha iyi sonuçlar verir, ancak sınıflandırıcının oluşturulması hesaplama maliyetini artırır. Son olarak, gömülü yöntemler, FS işlemini gerçekleştirmek için sınıflandırma modelinin dahili bilgilerini kullanır (ör. SVM’de ağırlık vektörünün kullanımı). Hesaplama maliyetlerinin sarmalayıcı yöntemlerden düşük olması ve sınıflandırma performanslarının filtreleme yöntemlerine göre üstün olması sebebiyle genellikle performans ve hesaplama maliyeti arasında iyi bir denge sağlar.

Biyobelirteç tespiti için düşük maliyetli ve kolay uygulanabilir olması sebebiyle en

yaygın olarak tercih edilen yaklaşımlardan olan filtreleme yöntemlerinde, her OTU'ya örnek bolluk seviyeleri ile örnek etiketleri arasındaki uygunluğa dayalı bir puan atanır. En yüksek puanlara sahip OTU'lar potansiyel biyobelirteçler olarak ilan edilir. Bu puanlama işlemi, her OTU için tek tek ve diğer OTU'lardan ayrı olarak gerçekleştirilir. Bu nedenle, filtreleme yöntemleri hesaplama açısından hızlıdır ve kolay yorumlanabilir, ancak bireysel sıralama, farklı değişkenler arasındaki karşılıklı bağımlılıkları göz ardı eder. Bireysel sıralamanın aksine, öznelik dönüşümü tabanlı yöntemler, yeni tespit edilen her özneliğin, orijinal veri kümesindeki tüm özneliklere bağlı olduğu daha açıklayıcı öznelikler oluşturmaya çalışır. Yeni özneliklerin inşasında tüm başlangıç özneliklerini göz önünde bulundurmak, OTU'lar arasındaki etkileşimleri açıklamaya yardımcı olur. Dönüşüm yaklaşımları, örnek etiketlerinin dönüşüm sürecinde dikkate alınıp alınmadığına göre genel olarak denetimli ve denetimsiz olmak üzere iki kategoriye ayrılır. En çok kullanılan denetimli yaklaşımlara örnek olarak doğrusal ayırım analizi (Linear discriminant analysis, LDA) ve kısmi en küçük kareler (Partial least-squares) yöntemleri verilebilir, PCA ise en popüler denetimsiz yöntemlerden biridir.

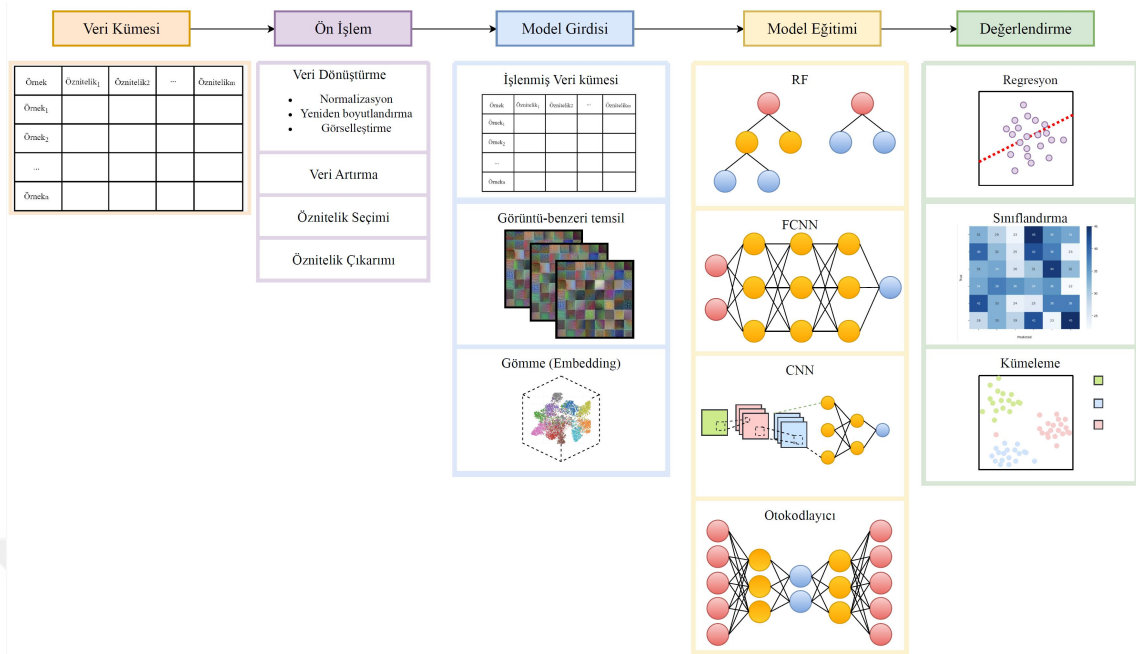
İstatistiksel testlere dayalı seçicilerin bir dezavantajı, çok sayıda ilişkili özneliği seçme eğilimleridir; birden fazla öznelik ilişkili ve ayırt edici ise, gerekli olmamalarına rağmen hepsinin seçilmesi muhtemeldir [63]. LDA, özneliklerin etki büyüklüğünü tahmin etmek ve bunları sıralamak için kullanılır, sınıfları hedef olarak koruyarak ve orijinal öznelikleri yardımcı değişkenlerle birlikte girdi değişkenleri olarak kullanarak eğitilir. Etki büyüklüğü, birinci doğrusal ayırmacı eksen boyunca sınıf ortalamalarındaki farkla niceliksel olarak belirlenir. LDA Etki Büyüklüğü (LEfSe), metagenomik verilerde FS için geliştirilen bir diğer popüler yöntemdir, istatistiksel anlamlılığın yanı sıra, biyolojik önemi de hesaba katar. RF ve GB Ağaçlar gibi ağaç tabanlı modeller, giriş özneliklerine bir önem puanı atar ve önem derecesine göre sıralanan öznelikler, son sınıflandırmadan sorumlu olan daha güçlü bir alt akış modeli için giriş verisi olarak kullanılabilir, ancak bu vekil modeller tarafından seçilen öznelikler, farklı kısıtlamalar ve optimizasyon yaklaşımları altında çalıştıkları için öngörücü için en uygun olanlar olmayabilir. Sarmalayıcı ve gömülü FS yöntemleri öngörücü modele bağlıdır [71]. En etkili FS yaklaşımı, veri kümesinin özelliklerine, yani seyrekliğine, özellik korelasyonuna ve kullanılan modele bağlıdır. Literatürde veriden bağımsız çalışan bir biyobelirteç keşfi

algoritması geliştirebilmek amacıyla çok sayıda çalışma yapılmıştır [1, 5, 12, 72], fakat halen tutarlı sonuçlar veren kararlı biyobelirteç bulma yaklaşımları araştırılmaktadır.

1.2.2. Mikrobiyom ve Derin Öğrenme

Derin sinir ağı olarak da bilinen DL, son yıllarda hesaplama gücünün artması, model mimarisindeki iyileşme ve verilerin katlanarak büyümesi nedeniyle büyük atılımlar yapan bir ML dalıdır. DL modelleri, insan beynindeki nöronları taklit eden çok sayıda yapay nöron katmanından oluşur. LR'ye benzer şekilde, her nöronun global kayıp fonksiyonunu en aza indirmek için geri yayılım sırasında gradyan iniş algoritması ile güncellenen bir ağırlık değeri vardır. Sigmoid, tanh veya relu gibi bir aktivasyon fonksiyonu kullanarak, girdi verileriyle çıktılar arasında türevlenebilir bir fonksiyon olarak tanımlanabilen bir ilişki kurar. Bu nedenle iyi eğitilmiş bir model, yeni, etiketlenmemiş verileri tahmin etmek için kullanılabilir. DL, ML'nin bir alt dalı olması sebebiyle temel olasılık ve istatistikler, kayıp / maliyet işlevi vb. dahil olmak üzere ML'deki bazı ortak bilgi temellerini devralır, ek olarak daha fazla esneklik ve daha iyi tahmin gücü için daha karmaşık katmanlara sahiptir. DL, belirli miktarda veriyle bir performans eşiğine ulaşan diğer birçok ML algoritmasını geride bırakarak büyük veri kümelerindeki üstün performansı ile öne çıkar. Dahası, DL teknikleri, genellikle insan gözlemiyle algılanamayan karmaşık özellikleri ortaya çıkarmak için güçlü bir kapasiteye sahiptir ve buna 'temsili öğrenimi' denir [58]. DL'nin, görüntü tanıma, metin işleme ve otomatik çeviri gibi çok çeşitli alanlardaki yüksek başarısı nedeniyle, giderek daha fazla araştırmacı mikrobiyom veri analizinde DL tekniklerini kullanmaya başlamıştır [73, 74]. DL'nin metagenomik alanında temel kullanım amaçları özetle sınıflandırma, kümeleme ve boyut indirgeme olarak kategorize edilebilir.

ML ve DL yöntemleri, mikrobiyom veri analizinde iki farklı aşamada kullanılabilir; veri ön işleme (öznitelik seçimi gibi) ve model eğitimi (denetimli öğrenme gibi). Veri analizinde çeşitli görevler için ML/DL yöntemlerinin kullanımına dair genel süreç, Şekil 1.3 'de gösterilmiştir. Veri kümeleri doğrudan bir modelin girdisi olarak kullanılabilir veya ön işlenebilir. Ön işleme adımı genellikle modelin tahmin sonuçlarını iyileştirir. Örneğin model, mikrobiyomun hiyerarşik ilişkisini önceden öğrenebilirse daha iyi bir tahmin modeli oluşturmak daha kolay olacaktır [75]. DL, diğer yöntemlerle



Şekil 1.3. ML/DL yöntemlerine dair kapsamlı süreci içeren genel bir akış şeması.

karşılaştırıldığında, ayrık veya sürekli verilere uygulanmak için daha esnek ve geneldir ayrıca ML'ye kıyasla, genel olarak daha az uzmanlık bilgisi ile daha az mühendislik özelliği gerektirir. Giriş verileri; metin verileri, mikrobiyom verileri, ifade verileri (yani, mRNA veya miRNA ifade verileri), genomik dizi verileri, sayısal veya kategorik veriler olabilir. Yüksek boyutlu metagenomik verilerde boyut indirgeme isteğe bağlı bir adımdır. Öznitelikler daha sonra bir DL/sinir ağı modeli oluşturmak için kullanılır. Kullanılacak modellerin türü, giriş verilerine bağlıdır. Örneğin, görüntü verilerinde genellikle evrişimli sinir ağı (CNN) modeli kullanılır. Birden fazla veri türü mevcutsa hibrit modeller oluşturulabilir. Model, eğitimi tamamlandıktan sonra, test (veya doğrulama) veri kümeleri ile test edilir. Değerlendirme için veri kümelerini kullanarak modelleri test etmek ve karşılaştırmak da oldukça önemlidir. İşlemler sonunda model uygulamalarda kullanılabilir.

1.2.2.1. Mikrobiyom Verilerinde Kullanılan Derin Öğrenme Yöntemleri

Çeşitli tipteki sinir ağları (NN) metagenomikte oldukça geniş bir uygulama alanına sahiptir. Bunlardan ilki olan, en basit NN yapısına sahip geleneksel ileri beslemeli sinir ağı (FFN), tek yönlüdür ve tekrarlayan bağlantıları yoktur bu nedenle, bilgi yalnızca bir yönde (giriş düğümlerinden çıkış düğümlerine) ve gizli düğümler (eğer ağda gizli

düğümmler varsa) aracılığıyla akar. Daha sonra hata, düğümler arasındaki bağlantılar aracılığıyla geri yayılır ve ağıın performansını iyileştirmesine olanak tanır. FFP'ler basit bir mimari olarak kabul edilir ve bu tür ağılar karmaşık verilerle ilgili olarak ortaya çıkabilen aşırı uyum ve gradyanların kaybolması (vanishing gradients) gibi problemlerle karşı karşıyadır, bu sebeple genellikle diğer ağıların bir alt kümesi olarak kullanılır [58]. Bir diğer temel sinir ağı mimarisi, bir katmandaki düğümlerin sonraki katmandaki tüm düğümlere bağlı olduğu tam bağlantılı sinir ağıdır (FCNN). MetaNN yönteminde [76], bu mimari ham metagenomik verilerinden konak fenotipini tahmin etmek için kullanılmış ve farklı veri kümeleri arasında geleneksel yöntemlere göre daha iyi sınıflandırma doğruluğu elde etmiştir. FCNN etkili bir bağımsız model olsa da çoğu zaman çok daha karmaşık mimarilerin temel yapı taşı olarak kullanılır. Çok Katmanlı Algılayıcı (MLP), ardışık katmanların birbirine tamamen bağlı olduğu bir NN'dir. MLP'leri kullanan dikkate değer bir yaklaşım, filogenetik bir ağaca dahil edilen göreceli bolluklarla çalışan bir model olan DeepEn-Phy'dir [77]. Bu modelde filogenetik ağaçlar, yapraklardan köklere doğru seviyelere ayrılır ve her bir düğümden bir MLP yer alır. İlk MLP'ler yaprak düğümlerindeki göreceli bollukları kullanır ve çıktılarını, kök MLP'ye ulaşana kadar diğer MLP'lere beslenir. Bu model, bağırsak mikrobiyomunu kullanarak sigara içme durumu ve vücut kitle indeksi tahmininde, RF, GB ağaçlar ve filogenetik ağaçlar üzerinde çalışan bir diğer DL modeli olan PopPhy-CNN'den daha iyi sonuçlar elde etmiştir.

CNN'ler özellikle görüntü sınıflandırmadaki yüksek performanslarıyla bilinmektedir. Omurgalıların korteksinden esinlenerek, veriler arasında hareket eden ve yerel mekânsal öznitelikleri çıkaran evrişim işlemine dayalı filtreler kullanan CNN'ler, metagenomik alanında ortak nükleotid desenleri gibi ortak yerel desenlere sahip dizileri sınıflandırmak ve aynı zamanda mikrobiyomun yapısını karakterize etmek için kullanılmaktadır [58]. Araştırmacılar, OTU bolluk matrislerini mekânsal bilgilerle (örneğin filogenetik ağaçlar) zenginleştirmek için yaratıcı yollar aramıştır. Bu gayeyle, tümevarımsal yetenekleri dolayısıyla en uygun araçlardan biri olan CNN'lerden çokça yararlanılmıştır. CNN'ler giriş verilerindeki yerel yapıyı özetlemek konusundaki üstün becerileri sebebiyle, görüntüler gibi mekânsal bilgi ileten verileri işlemek için oldukça uygundur. Nguyen vd. [78], her bir örneği bir kareye yeniden şekillendirerek ve ayrıca her pikseli mikrobiyal taksonların bolluğuna veya varlığına göre renklendirerek, OTU tablosunu bir görüntüye

dönüştürmüşlerdir. Bir diğer yöntem olan PopPhy-CNN [67]'de ise OTU bolluk verileri kullanılarak bir filogenetik ağaç oluşturulmuş ve ardından ağaç iki boyutlu bir matrise dönüştürülmüştür. Genel olarak, bu yöntemler farklı yaklaşımlarıyla konak fenotipi tahmini görevinde kıyas alınan yöntemleri (hem geleneksel ML yöntemleri hem de FCNN'ler) geride bırakmıştır.

Yinelemeli sinir ağları (RNN) bilgiyi bir noktadan diğerine geçiren zincir benzeri döngülerden oluşur ve çoğunlukla ardışık kalıpları keşfetmek için kullanılır. RNN'ler, mikrobiyom çalışmalarında zamansal bağımlılıkların ve dinamik kalıpların tahmin edilmesine olanak tanır. Örneğin Metwally vd. [79], boylamsal mikrobiyom profillerine dayalı bir tahmin modeli oluşturan öncü çalışmaları ile bebeklerin alerjik fenotipini üç yıl boyunca izleyen bir çalışmadan elde edilen verilerle, gıda alerjisini üstün bir şekilde tahmin etmeyi başarmıştır. Benzer şekilde, RNN tabanlı bir çerçeve olan phyLoLSTM [80] ise öz nitelik çıkarımı için taxoNN [81] kullanarak önceki sınıflandırma doğruluğunu iyileştirmiştir.

1.3. Derin Öğrenme ile Mikrobiyom Analizinde Karşılaşılan Problemler

Mikrobiyomun insan konak üzerindeki güçlü etkisi nedeniyle, mikrobiyom bileşimindeki değişikliklerin, konakta nasıl fizyolojik değişikliklere yol açabileceği konusu giderek daha fazla önem kazanmaktadır. Mikrobiyom çalışmalarında, mikroplar ve konağın sağlık durumu arasındaki temel ilişkiyi ortaya çıkarabilmek, önemli analizler gerektirir. Biyolojik veriler karmaşık ve hiyerarşiktir, basit analiz araçlarıyla değerli bilgiler elde etmek kolay değildir. DL yaklaşımları, desenleri otomatik olarak keşfetme konusunda, geleneksel ML yöntemlerine kıyasla çeşitli avantajlar sağlar. Ancak, DL tabanlı mikrobiyal veri analizi için hala birçok zorluk bulunmaktadır [75].

Mikrobiyom verilerinin karmaşıklığı

Metagenomik veriler karmaşıktır ve zorluklar çeşitli düzeylerde ortaya çıkabilir. İlk olarak, ham verilerin oluşturulması karmaşık ve pahalı (ancak giderek daha az pahalı) ıslak laboratuvar protokolleri gerektirir. Örnekler, doğal veya konakçı olsun, köken ortamlarından çıkarılmalıdır. Tercih edilen dizileme teknolojisi, örnek toplama yöntemleri, DNA çıkarma kitleri ve kütüphane hazırlama protokolleri ile birlikte,

çıkarılan mikrobiyom profilinin gerçek mikrobiyom bileşimini ne kadar iyi yansıttığını etkileyebilecek sistemik yanlılıklara yatkındır. Ayrıca metagenomik dizileme verilerinin biyoinformatik işlenmesi ve kantitatif metagenomik profilleri oluşturmak için kullanılan referans veri tabanları, profillerin gerçek mikrobiyom bileşimine göre doğruluğunda önemli bir rol oynar, yani incelenen mikrobiyal topluluğun üyeleri tercih edilen veri tabanında iyi temsil edilmiyorsa, bu durum sistemik yanlılıklara neden olabilir [58]. Son olarak, farklı veri dönüşümleri, metagenomik özelliklerin mutlak bolluğunu tahmin etmek için farklı yaklaşımların kullanımı veya dizileme derinliğindeki varyasyonları ele almak için benimsenebilecek farklı stratejiler vasıtasıyla, metagenomik verilerin kompozisyonellik açısından özel doğasını ele almak için yapılan seçimler, daha sonra biyobelirteç keşfi veya fenotip tahmini için kullanılacak niceliksel profiller üzerinde büyük bir etkiye sahiptir [58]. Özetle, mikrobiyom verilerinin içsel karmaşıklığı ve değişkenliği, doğru ve tutarlı hastalık sınıflandırması için önemli zorluklar sunmaktadır. Tüm bu protokoller, tıbbi metagenomik örneklerin gizlilik endişeleriyle birlikte, bu verileri hala nadir ve pahalı hale getirmektedir.

Yüksek Boyutluluk

Metagenomiğin doğası gereği, genellikle yüksek boyutlara ve buna karşılık nispeten daha az sayıda örneğe sahip veri kümeleri üretilmektedir. Bu kombinasyon, DL algoritmalarının karar süreçlerinde kullanılabilecek kalıpları bulmasını zorlaştırır. DL algoritmaları, birçok parametreyi eğiterek çalışır ve klasik algoritmalarla aynı düzeyde öznitelik seçimine ihtiyaç duymaz. Uygun veri gösterimlerini kendi başına çıkarabilme yeteneği, en güçlü yanlarından biri ve tanımlayıcı bir özelliğidir ancak bu yetenek, algoritmanın karmaşık desenleri tanıyabilmesi için genellikle büyük veri kümelerine erişim gerektirir [57]. Fakat hasta verileri söz konusu olduğunda büyük veri kümeleri elde etmek zor olabilir, çok sayıda numuneyi toplama ve işleme süreci hala pahalıdır ve yasal gerekliliklerin de hesaba katılması gerekir.

Büyük veri genellikle hacim, hız ve çeşitlilik ile tanımlanır. Bunların her biri metagenomikte önemlidir ve farklı bir zorluğu temsil eder. Veri hacmi, metagenomik örneklerde bulunan mikroorganizmaları tanımlamak ve ayırt etmek amacıyla çok sayıda dizileme okumasından kaynaklanır. Tek bir çalışma için yüklü miktarda veri ortaya çıkar,

hatta veri işleme sırasında (örneğin, metagenomik birleştirme için), gerekli sistem belleği birkaç terabayta ulaşabilir. Tüm ara ve nihai sonuçların güvenli bir şekilde arşivlenmesi ve depolanması, ihtiyaç duyulan depolama miktarını kat kat artırması nedeniyle başka bir zorluk sunar [82]. Ayrıca arşivleme, minimum ve maksimum saklama süreleri dahil olmak üzere tüm yasal gerekliliklere uymalı ve çeşitli veri türleri için otomatik bir şekilde yapılmalıdır.

Örnek Sayısının Yetersizliği

DL modellerini eğitmek ve farklı veriler için genelleştirebilmek, yüksek örnek sayısına sahip veri kümeleri gerektirmektedir. Bugüne kadar literatürdeki modellerin çoğunu eğitmek için kullanılan metagenomik örnek sayısı, yalnızca yüzlerce, en iyi ihtimalle binlercedir ve bu miktar DL modellerini verimli bir şekilde eğitmek için gereken miktarın altındadır. Aynı şekilde bu veri kümelerinin çeşitliliğinin yetersiz olması, yine farklı sorunlara işaret etmektedir. Çünkü bağırsak mikrobiyomu bileşimi diyet, yaş, genetik ve çevre dahil olmak üzere çok sayıda faktörden etkilenir. Bir model, spesifik bir hastalığı tespit etmek için tasarlanmış ve yalnızca belirli bir türdeki hastalardan alınan verilerle eğitilmişse, farklı kökenlere veya yaşam tarzlarına sahip hastalar için genelleştirilemeyebilir [83]. Bu nedenle, çeşitli popülasyonlardan büyük veri kümeleri kullanmak, daha genelleştirilebilir modellerin ve dolayısıyla daha fazla sayıda hasta için kullanılabilecek yaklaşımların önünü açabilir.

Kara Kutu

NN'lerin uzun süredir devam eden problemlere uygulanması, özellikle son on yılda performans ve tahmin gücünde önemli bir atılım sağlamıştır. Ancak artan model karmaşıklığından kaynaklanan yüksek performans, genellikle yorumlanabilirliğin kaybı pahasına gelir, yani birçok derin model kara kutu gibi davranır ve tahminleri hakkında açıklama sağlamada başarısız olur. Bu durum belirli uygulama alanları için ikinci plana atılabilir, fakat yüksek riskli alanlarda, örneğin sağlık hizmetlerinde, bir modele güven oluşturmak ve davranışını anlayabilmek çok önemlidir. Yorumlanabilir bir modelin, yalnızca mikrobiyal veri analizi için değil, biyoinformatik için de hayati önem taşımasının iki önemli nedeni olduğuna inanılmaktadır. İlk olarak, model eğitilirken tahminlerin nasıl yapıldığını anlamak, girdi verilerindeki hataları veya yanlılıkları belirlemek için

önemlidir. İkinci olarak ise model ancak yorumlanabiliyorsa büyük biyolojik verilerden anlamlı içgörülere rehberlik edecek olan yeni desenler öğrenebilir [75]. Özetle, DL'nin yorumlanabilirliği için bütünleşik ve standart bir tanımın olmaması, çoğu durumda çalışmaların evrenselliğinin önüne geçmektedir.

1.4. Araştırmanın Amacı ve Önemi

Dizileme teknolojisinin günümüzdeki karşılanabilirliği, çeşitli sistemler üzerinde kapsamlı araştırmalara olanak tanır ve bunun sonucunda yüzlerce, hatta binlerce örnek için çok sayıda mikrobiyal veri üretilir. Bu büyük miktarda verinin yorumlanabilmesi için, kapsamlı veri kümelerinden önemli içgörülerini etkili bir şekilde çıkarabilen metodolojilere ihtiyaç duyulur [15]. Algoritmaların bu yüksek boyutlu verilere uygulanması, boyutluluğun laneti olarak bilinen önemli bir zorluğa yol açar. Bu zorluğun iki yönü vardır: birincisi, çok sayıda öznitelik, hesaplama maliyetini artırır, ikincisi, nispeten düşük sayıda örnek diğer veri kümelerine genelleştirilebilirliğin azalmasına yol açar. Bu tez çalışmasında, bağırsak mikrobiyom verilerinin yüksek boyut-az örnek sorununu hafifletebilmek amacıyla, verileri daha düşük boyutlu öznitelik uzaylarında modelleme girişimleri temel bir temsili öğrenme yaklaşımı olarak benimsenmiştir.

Diğer bir husus olan mikrobiyom verilerinin içsel karmaşıklığı ve değişkenliği, doğru ve tutarlı hastalık sınıflandırması için önemli zorluklar sunmaktadır. Büyük kohort çalışmaları için dikkate değer miktarda mikrobiyom örneğinin toplanması, daha önce ağırlıklı olarak büyük ve çok merkezli projeler tarafından gerçekleştirilmiştir; bunun başlıca nedeni, yüksek maliyetler ve laboratuvarlar için altyapı eksikliğidir. Ancak, kapsamlı veri paylaşım politikalarının teşvik edilmesindeki son gelişmeler ışığında, analizlerin daha önce yayınlanmış çalışmalardan elde edilen veriler bir araya getirilerek veya yeniden kullanılarak yürütülmesi, mikrobiyom araştırmalarında yeni hipotezlerin üretilmesinde veya doğrulanmasında kritik roller üstlenmeye başlamıştır. Çoğu metagenomik veri, artık dünya çapında yönetilen birden fazla bağımsız araştırma sayesinde erişilebilir olduğundan, bu veri kümelerinin birleştirilmesi, sağlık ve hastalıkla ilişkili imzaları büyük ölçekte incelemek ve sınırlı veya bireysel çalışmalarla elde edilemeyen yeni, kapsamlı içgörüler elde etmek için cesaretlendirici bir strateji sağlayabilir [84]. Ayrıca, her biri farklı hastalık türlerine odaklanan farklı hastalık

çalışmalarından veri kümelerinin birleştirilmesi, hastalıkla ilgili daha geniş bir veri yelpazesi sağlar ve verilerdeki bu çeşitlilik, izole çalışmalarda tespit edilemeyebilecek, daha genel bir hastalık-sağlık örüntüsü veya imzası ortaya çıkarabilir. Sonuç olarak, bu yaklaşım çeşitli hastalık durumlarından elde edilen toplu içgörülerden yararlanarak daha kapsamlı bir sağlık değerlendirmesine olanak sağlayabilir. Bu sebeple, bu tez çalışmasında İBH ve KK mikrobiyom verilerini içeren dokuz ayrı veri kümesi birleştirilmiş ve geliştirilen DL algoritmaları bu geniş çaplı veriler üzerinde test edilmiştir. Bunun yanında, geliştirilen modellerin sınıflandırma doğruluğu, sağlamlığı ve genelleştirilebilirliği ile ilgili analizler, yine benzer şekilde birleştirilmiş on bir farklı hastalıktan mikrobiyom verileri içeren kapsamlı bir veri kümesi ile yapılmıştır.

Biyobelirteçler, biyolojik süreçleri, hastalık ilerlemesini veya tedaviye yanıtı değerlendirmek için kullanılabilen ölçülebilir göstergelerdir ve hastalık teşhisi, prognozu ve tedavi takibinde önemli bir rol oynarlar [1]. Metagenomik verileri kullanarak mikrobiyom çapında ilişki çalışmaları yürütmek, hastalıklarla ilişkili mikrobiyal biyobelirteçlerin tanımlanmasında etkilidir. Biyobelirteçlerin tespiti, metagenomik verilerin klinik uygulamalara entegre edilmesi için önemli bir araçtır. Metagenomik verilerden hastalıkla ilgili biyobelirteçleri tanımlama süreci, sağlık ve hastalık veya çeşitli hastalık evreleri gibi iki veya daha fazla fenotip arasında ayırım yapabilen etkili bir biyobelirteç alt kümesi bulmayı içerir [85]. Ancak bu süreç, pratikte farklı sebeplerden dolayı zorlaşabilir. Örneğin, tek yumurta ikizlerinin bağırsak bakteri ekosistemlerindeki önemli farklılıklarla kanıtlandığı gibi [86], çok sayıda mikrobiyal popülasyon kişiden kişiye önemli derecede değişkenlik göstermektedir. Özellikle bu alanlar için FS kullanırken, yalnızca model performansı değil, aynı zamanda FS sürecinin gürbüzlüğü de önemlidir, çünkü alan uzmanları, veri kümesindeki yalnızca küçük değişiklikler için, kararsız olana göre kararlı bir FS algoritmasını tercih ederler. Çoğu durumda seçilen potansiyel öznitelikler daha sonra analiz edilir ve özellikle biyomedikal uygulamalarda bu işlem çok zaman ve çaba gerektirir. Bu sebeple gürbüz FS teknikleri, alan uzmanlarının seçilen özniteliklere daha fazla güvenmesine olanak tanıyacaktır.

Özellikle son on yılda, hastalığın gerçek ilerlemesini yansıtacak ve dolayısıyla İBH ve KK gibi hastalıkların yönetiminde önemli bir rol oynayacak biyobelirteçlerin araştırılması için kapsamlı araştırmalar yapılmıştır. İdeal bir invaziv olmayan biyobelirteç

tespit algoritması, kolay erişilebilir, tekrarlanabilir, ekonomik, oldukça spesifik ve hassas olmalıdır. Şu anda, KK'nin erken teşhisi için kolonoskopi, X-ışını baryumlu lavman, dışkıda gizli kan testleri ve tümör belirteci tespiti gibi bazı klinik yaklaşımlar bulunmaktadır [87]. Kolonoskopi ve X-ışını baryumlu lavman, erken KK tespiti için klinik önem taşır ve kesin tanı için çok önemli olsa da, genel popülasyonda yaygın uygulamaları yüksek maliyetler ve invazivlik nedeniyle sınırlıdır [26]. Sonuç olarak, erken KK taramasını geliştirmek ve KK'li bireyler için tedavi potansiyelini artırmak için yüksek duyarlılık ve özgüllüğe sahip invaziv olmayan biyobelirteçleri tespit etme ihtiyacı devam etmektedir [87]. Biyolojik verilerden elde edilen sonuçların tekrarlanabilirliğinin sağlanması, klinik uygulamalar ve hatalı biyolojik bulguların önlenmesi için büyük önem taşımaktadır, ancak yukarıda bahsedilen zorluklar nedeniyle, hastalık tespiti için gürbüz biyobelirteçler bulmak zorlaşmaktadır. Bu nedenle bu çalışmada İBH ve KK hastalıklarına yönelik, örnek sayısının az olmasına rağmen yüksek sınıflandırma performansı sergileyen modellerin geliştirilmesi ve erken tanı için kullanılabilecek gürbüz biyobelirteçlerin belirlenmesi amaçlanmıştır.

Bu tezde yapılan çalışmalar kategorik olarak ikiye ayrılabilir. İlk olarak, mikrobiyom verilerinde sınıflandırma doğruluğunun artırılması amacıyla iki farklı yaklaşım geliştirilmiştir. DeepGum isimli ilk yaklaşım, yüksek boyutlu verilerin getirdiği önemli zorluğun üstesinden gelmek ve düşük örneklemlili mikrobiyom veri kümelerinde daha yüksek doğrulukta sonuçlar elde etmek için, boyutsallığı azaltmanın tek bir veri kümesinde geleneksel kullanımının aksine, darboğaz modellerine, metagenomların çok kohortlu bir koleksiyonunda temsili öğrenenler olarak yaklaşmayı önerir. İkinci yaklaşım olan Deeptaxim ise bağırsak mikrobiyomu değişiklikleriyle güçlü ilişkileri olduğu bilinen hastalıklara odaklanır ve mikrobiyom verilerini taksonomik ağacın hiyerarşik yapısını kullanarak görüntü verisi olarak işler. Bu yaklaşımın temel amacı, mikrobiyom verilerini görüntü verilerine dönüştürerek ve bunları CNN tabanlı yöntemlerle eğiterek daha fazla içgörü elde etmek, mikrobiyom tabanlı hastalık sınıflandırmasının doğru bir şekilde yapılması için DL yöntemleri geliştirmek ve optimize etmektir. İkinci kategoride ise bir biyobelirteç keşif çalışması olan Disrobiom metodu bulunmaktadır. Bu yaklaşım farklı veri kümelerinde aynı hastalıklara ait tutarlı ve gürbüz biyobelirteçleri bulabilmeyi amaçlar ve bu amaçla yaklaşım çeşitli kararlılık testlerine tabii tutulur.

2. BÖLÜM

ARAŞTIRMA YÖNTEMLERİ

2.1. Gürbüz Temel Bileşen Analizi

Gürbüz temel bileşen analizi (Robust principal component analysis, RPCA) yöntemi [88], ilk kez Candes E. J. vd. tarafından, girdilerinin bir kısmı keyfi olarak bozulmuş olsa bile, bir veri matrisinin temel bileşenlerinin kurtarılabileceği iddiasıyla öne sürülmüştür. Dolayısıyla, bu yaklaşımın düşük dereceli matrisleri ve yüksek boyutlu alanlardaki seyrek sinyalleri kurtarmadaki yeteneğinin, görüntü/video işleme, web veri analizi ve biyoinformatikte ortaya çıkan büyük veri kümelerini işlemek için güçlü bir araç olabileceği düşünülmüştür. Bu motivasyonla, Alshawaqfeh vd. [1] metagenomik bir veri kümesinde, ilgili biyobelirteçleri tespit etmek amacıyla RPCA yöntemini kullanmıştır. Motivasyonun temeli, mikropların çoğunluğunun eldeki biyolojik süreçle alakasız olduğu gerçeğine dayanmaktadır. Bu alakasız özneliliklerin, iki farklı fenotip (yani sağlıklı ve hasta) arasında değişmeyen bolluk seviyelerine sahip olduğu varsayılmaktadır. RPCA yöntemi, orijinal veri matrisini (D), seyrek bir matris (S) ve düşük dereceli bir matris (L) olacak şekilde iki matris olarak modeller. Bolluk düzeyi, düşük dereceli bir matris olarak düşünülebilir. Öte yandan, ilgili biyobelirteçlerin bolluk seviyeleri iki fenotip arasında önemli farklılıklar gösterir, bu da seyrek bir matrisle temsil edilebilir. RPCA, giriş verilerini içeren orijinal matrisi, L ve S matrislerinin süperpozisyonuna ayırmak için kullanılır ($D = L + S$). Ardından, S matrisine dayalı potansiyel biyobelirteçler belirlenir. RPCA, Temel bileşen takibi (Principal Component Pursuit, PCP) olarak adlandırılan bir optimizasyon problemini çözerek, hem L hem de S matrisini bulabilme özelliğine sahiptir. PCP, düşük dereceli matrisin nükleer normunun ve seyrek matrisin L_1 normunun ağırlıklı toplamını en aza indirmeyi amaçlar:

$$||L||_* + \lambda ||S||_1 \quad (2.1)$$

PCP probleminin çözümü için Artırımlı Lagrange Çarpımı (Augmented Lagrange Multiplier, ALM) kullanılır. ALM algoritması, artırımlı lagrange adı verilen, yeni bir amaç ile sınırlandırılmış optimizasyon problemlerini kısıtsız problemlere dönüştürerek çözer. λ , pozitif düzenleme parametresi, Y , Lagrange çarpan matrisi ve μ , ALM formülasyonu ile ilişkili düzenleme parametresi olmak üzere PCP'nin ALM ile formülasyonu denklem 2.2'de verilmiştir.

$$l(L, S, Y) = ||L||_* + \lambda ||S||_1 + \langle Y, D - L - S \rangle + \frac{\mu}{2} ||D - L - S||_F^2 \quad (2.2)$$

Denklem 2.2'nin çözümü için yineleme tabanlı bir yöntem kullanılır ve her yineleme (k) adımdan oluşur ($(L_k^*, S_k^*) = \arg \min_{L, S} l(L, S, Y)$). S_k^* ve L_k^* için kapalı bir çözüm olmasına rağmen, alt problemi (denklem 2.2) çözmek, optimal çözüme yakınsayana kadar tekrar hesaplama (denklem 2.3 ve 2.4) gerektirir. Bu tekrar, önemli bir hesaplama yüküne sebep olur. [89]'e göre S ve L 'nin bir kez güncellenmesi yeterlidir.

$$S_k^* = \arg \min_S l(L, S, Y) = S_{\lambda \mu^{-1}}(D - L + \mu^{-1} Y) \quad (2.3)$$

$$L_k^* = \arg \min_L l(L, S, Y) = D_{\mu^{-1}}(D - S + \mu^{-1} Y) \quad (2.4)$$

Bu iterasyon sayısı alt problemin optimal çözümünü vermese de RPCA'nın optimal çözüme yaklaşması için yeterlidir. D matrisinde her sütun, bir örnekteki tüm mikropların diferansiyel bolluk seviyelerini ve her satır, bir mikrobun tüm numunelerdeki diferansiyel çeşitliliğini temsil eder. S 'nin girişleri, biyolojik sürece yanıt olarak pozitif veya negatif olabilir. Bu nedenle, diferansiyel olarak bol bakterilerin (yani biyobelirteçlerin) tanımlanması için S 'deki girişlerin mutlak değerlerine ihtiyaç vardır. i . bakterinin skoru, S matrisindeki i . satırın mutlak değerlerinin toplanmasıyla hesaplanır. Skor vektörü, S 'nin elemanlarının mutlak değerlerinin toplanmasıyla elde edilir. Büyük puanlar, iki durum arasında daha büyük çeşitlilik sergileyen mikroplarla ilişkilidir. Bu nedenle, biyobelirteç

olarak sadece en yüksek skora sahip taksonlar seçilir.

2.2. Yapay Sinir Ağları

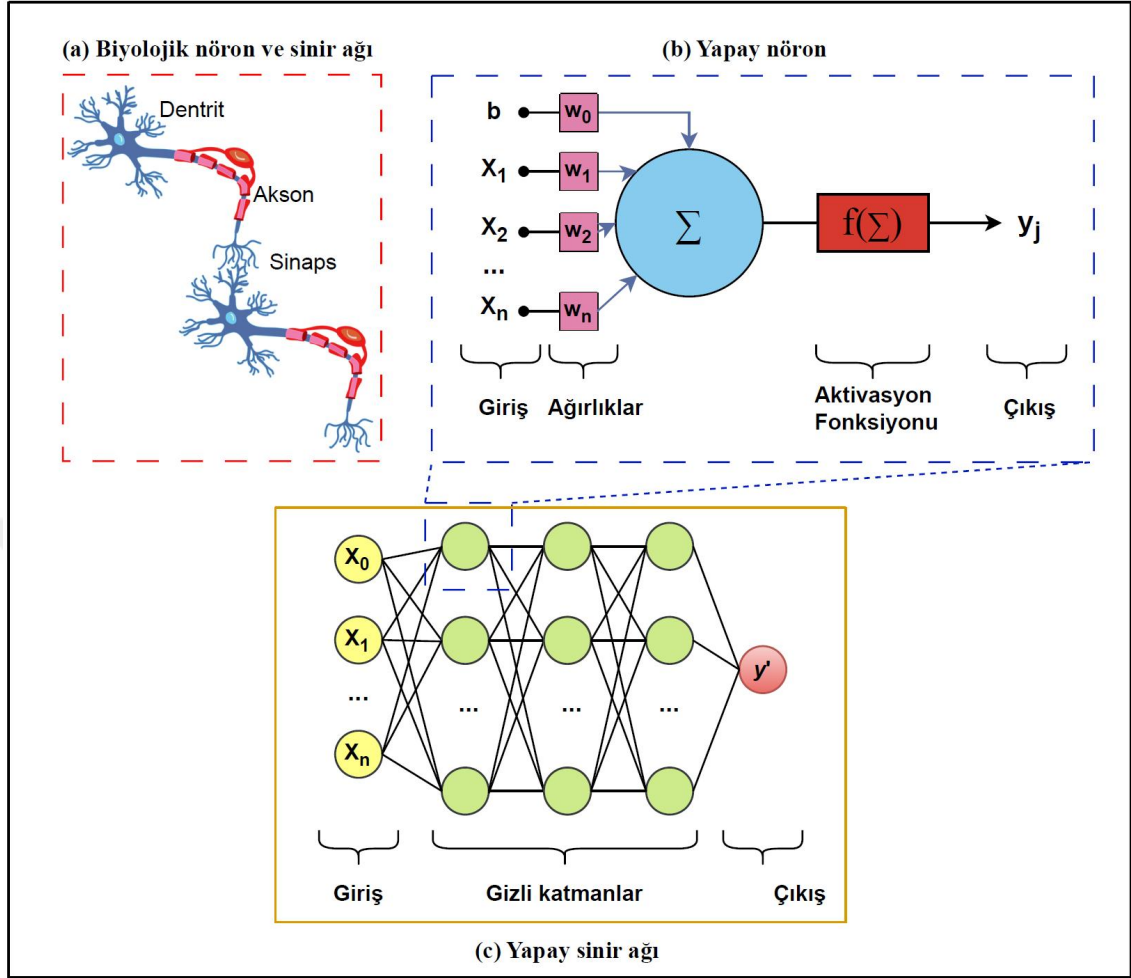
Yapay sinir ağları (YSA), insan beynini modelleme çabasıyla oluşturulmuş, fakat seneler içerisinde ilk amaçlarından farklı olarak, birçok alanda güçlü ML araçları olarak kullanılmıştır. Her YSA, diğer düğümlerle bağlantılar (aksonlar ve dendritler) aracılığıyla iletişim kuran düğümler (hücre gövdelerine benzer) içerir. Biyolojik sinir ağında (Şekil 2.1-a), nöronlar arasındaki sinapsların, nöronları ilişkili çıktılara sahip olduğunda güçlendirmesi gibi, YSA'daki düğümler arasında bağlantılar da istenen bir sonucu sağlama yeteneklerine göre ağırlıklandırılır [90]. Sinir ağlarının yapısı, her biri çoklu nöron içeren çoklu katmanlara ayrılacak şekilde tasarlanabilir. Her bir nöron (j), Şekil 2.1-b' de gösterildiği gibi, $n = (0, 1, \dots, n-1)$ olmak üzere, n elemanlı bir X giriş vektörü, ağırlık parametreleri (w), bir bias (b) ve bir önceki katmanın çıktılarının ağırlıklı toplamını (z) girdi olarak alan bir aktivasyon fonksiyonu (f) içerir.

$$z_j = \sum_{n=0}^{N-1} w_{jn}x_n + b_j \quad (2.5)$$

$$y_j = f_j(z_j) \quad (2.6)$$

Ağırlıklı toplamın matematiksel ifadesi denklem 2.5' te verilmiştir. Her bir nöronun çıkış değeri (y_j) ise denklem 2.6'daki gibi hesaplanır ve üretilen her bir çıkış bir sonraki nöronun girişidir [91]. Hesaplamaların devamlı ileri yönde beslenmesi dolayısıyla bu yaklaşım ileri yayılım olarak adlandırılır. YSA'lar bu yaklaşımı, her biri çok sayıda nöron içeren katmanlar için de kullanmaktadır, her bir katmanın nöronları bir önceki ve sonraki katmanlara bağlıdır. Şekil 2.1-c'de $n = (0, 1, \dots, n-1)$ olmak üzere, her giriş (x_n) için bir nöron içeren bir giriş katmanı, bir önceki katmanın çıktısını girdi olarak alan gizli katmanlar ve çıkış katmanını gösteren, örnek bir YSA yapısı verilmiştir.

Her bir örnek (x_n) eğitim aşamasında, giriş katmanı üzerinden kendisine karşılık gelen y etiketi ile beslenir ve çıkış katmanı adı verilen son katmanın, çıkış etiketinin yaklaşık bir değeri olan $\hat{y}$ etiketini vermesini bekler. Sinir ağı, verilerin hiyerarşik



Şekil 2.1. a: Biyolojik nöron ve sinir ağının temsili gösterimi, b: Temel yapay nöron gösterimi, c: Örnek bir yapay sinir ağı.

bir temsili oluşturmak ve çıktı katmanında doğru tahminler elde edebilmek için ara aşamalarındaki temsillerin en iyi şekilde nasıl kullanılacağını öğrenebilmelidir. Bir YSA'yı eğitmek için, ağın çıktılarını optimize etmek amacıyla ağırlıkları güncellemek gerekir. Bu sayede, öğrenme algoritması tahmin edilen çıkış değerleri ($\hat{y}$) ile gerçek çıkış değerleri (y) arasındaki kaybı minimize edebilir ve uygun ağırlık (W) ve bias (b) parametrelerini bulabilir. Parametrelerinin eğitimi ve optimize edilmesi, kayıp fonksiyonu dikkate alınarak geri yayılım yoluyla yapılmaktadır [91]. Geriye yayılım işlemi için, maliyet fonksiyonu veya kayıp fonksiyonu olarak adlandırılan, çıktıların doğruluğunu ölçen bir fonksiyon üzerinde, gradyan iniş adı verilen bir optimizasyon algoritması kullanılmaktadır.

YSA eğitiminde, temel amaç olan kaybın minimizasyonu şu şekilde gerçekleşir: eğitim

verileri ağ üzerinden beslenir ve zincir kuralı ile her ağırlığa göre kayıp fonksiyonunun gradyanı hesaplanır, sonrasında gradyan iniş algoritması kullanılarak tüm ağırlıklar güncellenir, bu işlemler kaybın minimize edildiği ana dek devam eder.

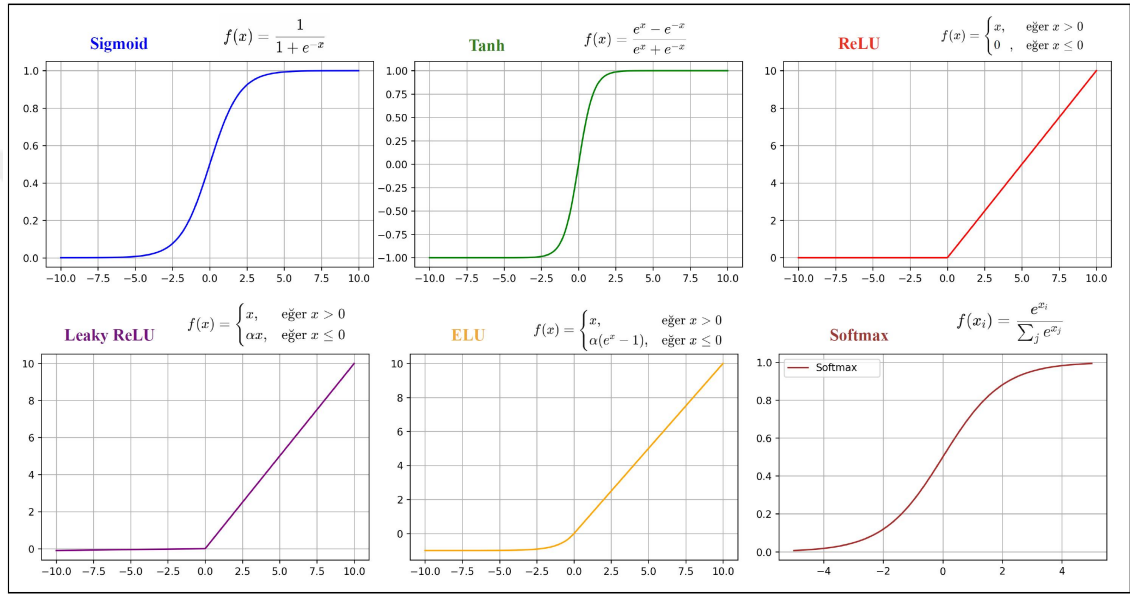
2.2.1. Aktivasyon Fonksiyonları

Aktivasyon fonksiyonları, bir katmanda bulunan nöronlardaki çıktı değerini sonraki katmanlara iletilirken kullanılır. Bu çıktı değerinin, diğer katmanlara iletilip iletilmeyeceğine karar verebilmek için bir eşik değeri belirlenmesi gerekir. Bir yapay sinir hücresindeki bilginin değeri $(-\infty, +\infty)$ aralığında olabilir ve nöron gerçek değerinin sınırlarını bilmeyebilir. Bu nedenle, nöronun aktif olup olmaması gerektiğine karar verebilmek için aktivasyon fonksiyonlarına ihtiyaç duyulur. Bu sayede, bir nöron tarafından üretilen çıktı değeri kontrol edebilir ve dış bağlantıların nöronu aktif olarak görüp görmeyeceğine karar verilebilir [92]. Genel olarak tüm derin ağlarda doğrusal bir dönüşümün ardından bir aktivasyon işlevi gelir. Aktivasyon fonksiyonu, derin sinir ağları (DSA) eğitiminin başarısında önemli bir rol oynar [93].

YSA'lar, daha çok doğrusal olmayan sınıflandırmalarda kullanıldığından, aktivasyon fonksiyonu genellikle doğrusal olmayan bir fonksiyon seçilir. Modelin öğrenme sürecinde geri yayılım algoritması için aktivasyon fonksiyonunun türevi de kullanıldığından, türevi kolay hesaplanabilen bir aktivasyon fonksiyonunun kullanılması algoritmanın hızı için önemlidir. Aktivasyon fonksiyonunun ağ mimarisi içerisindeki konumu, ağdaki işlevine bağlıdır. Aktivasyon fonksiyonu gizli katmanlardan sonra yerleştirildiğinde, öğrenilen doğrusal işlemleri, yayılım için doğrusal olmayan biçimlere dönüştürürken, çıktı katmanında kullanıldığında nihai tahminin hesaplanmasına yardımcı olur. Derin mimariler, verilen bir görevi çözmek için birlikte öğrenilen, her biri hem doğrusal hem de doğrusal olmayan işlemleri içeren birkaç işlem katmanından oluşur. Daha derin ağların daha iyi performanslar gösterdiği bilinmektedir ancak bu ağlarda gradyanların kaybolması ve genellikle 1' den küçük olan türev terimlerinin bir sonucu olarak gradyanların sonsuza gitmesi (exploding gradient) gibi ortak sorunlar ortaya çıkabilmektedir. Bu türev terimlerinin art arda çarpılmasıyla, değer küçülür ve sıfıra yönelir, böylece eğim kaybolur. Ayrıca değerler 1'den büyükse, art arda çarpma işlemi değerleri artıracak ve eğim sonsuza gidecektir. Bu sebeple, aktivasyon fonksiyonları

bu eğimlerin değerlerini belirli sınırlarda tutar. Bu fonksiyonlar, farklı matematiksel fonksiyonlar kullanılarak elde edilir.

DSA'lar ile yapılan çalışmalarda; kullanılan veri setinin yapısına, mimarinin boyutuna, derinliğine, kullanılan optimizasyon yöntemlerinin ve aktivasyon fonksiyonlarının çeşidine bağlı olarak model performanslarının farklılaştığı bilinmektedir. Şekil 2.2'de sıklıkla kullanılan aktivasyon fonksiyonları ve matematiksel formülleri verilmiştir.



Şekil 2.2. Sıklıkla kullanılan aktivasyon fonksiyonları ve matematiksel formülleri.

En standart aktivasyon fonksiyonlarından biri hiperbolik tanjant (tanh) fonksiyonudur. Tanh, orijin etrafında simetrik ve -1 ile 1 arasında değişen bir değer aralığına sahiptir. Tanh fonksiyonunun en büyük avantajı, sıfır merkezli bir çıktı üretmesidir, böylece geri yayılım sürecini destekler. Tanh'ın dezavantajlarından biri, büyük veya küçük girdiler için oluşan gradyanların doygunluğudur. Bu durum, ağların eğitimi yavaşlatabilir [94]. Sigmoid, tanh'a benzer şekilde DL mimarilerini oluşturmak için kullanılan en temel aktivasyon fonksiyonlarından biridir. Başlıca avantajlarından biri, 0 ile 1 arasında değişen bir değer aralığına sahip olmasıdır, bu nedenle özellikle bir olasılığı çıktı olarak tahmin etmeyi amaçlayan modeller için kullanılır. Sigmoid eğimi, çok büyük veya çok küçük girdiler için kaybolur ve bu da eğitim sürecini zorlaştırır. Ancak, gizli değişkenler içermeyen ve etiketlerine erişebilen çıkış birimleri için kullanılması durumunda sigmoid iyi bir seçenek olabilir [94].

ReLU, doğrusal olmayan fonksiyonların varsayılanı olarak kabul edilir. ReLU'nun temel avantajları; verimli hesaplama, önceki iki aktivasyon fonksiyonuna kıyasla daha az kaybolan gradyan problemi ve daha iyi gradyan yayılımı olarak sıralanabilir [94]. Krizhevsky vd., dört katmanlı bir evrişimli ağ için aktivasyon fonksiyonu olarak ReLU kullanıldığında gereken iterasyon sayısının, tanh fonksiyonuna göre altı kat daha az olduğunu belirtmiştir. Aktivasyon fonksiyonları arasındaki popüleritesi ve verimliliği sebebiyle, ilerleyen dönemlerde Leaky ReLU ve ELU [91] gibi birçok farklı varyasyonu önerilmiştir. Her iki varyasyon da benzer şekilde, bazı ReLU nöronlarının tüm girdiler için kullanılmadığı ve hangi girdi sağlanırsa sağlansın inaktif kaldığı ölmekte olan nöronlar (dying ReLU) sorununu ele almaktadır. Böyle bir durumda, bu nöronlardan hiçbir eğim akışı olmaz ve sinir ağı mimarisinin eğitimi etkilenir. Bu sebeple Leaky ReLU ve ELU, bir eğim ekleyerek ReLU'nun değer aralığını genişletir ve negatif değerlerin tamamının sıfır olmasını önler. Softmax aktivasyon fonksiyonları ise genellikle bir sinir ağının son aktivasyon fonksiyonu olarak kullanılır. Temel olarak, ağın çıktısını tahmin edilen çıktı sınıfları üzerinden bir olasılık dağılımına göre normalleştirir. Softmax fonksiyonu, her biri 0 ile 1 arasında bir değer depolayan ve olasılıklarının toplam değeri 1'e eşit olan N elemanlı bir vektör üretir. Softmax'ın sınırlamalarından biri, her giriş verisinin N sınıftan en azından birine ait olduğunu varsaymasıdır ancak pratikte durum böyle değildir, yani ağ, sınıflardan hiçbirine ait olmayan bir girdiye uygulanabilir.

2.2.2. Kayıp Fonksiyonları

Aktivasyon fonksiyonlarının yanında, kayıp fonksiyonu da maliyet fonksiyonunu tanımlayan sinir ağlarının temel unsurlarından biridir. Belirli bir tahmin için hatayı temsil eden fonksiyon olarak tanımlanır. Bu amaçla, belirli bir eğitim örneği için, tahmin edilen çıkışı gerçek çıkışla karşılaştırır. Sinir ağları için kayıp fonksiyonu neredeyse türevlenebilir herhangi bir fonksiyon olabilir [94]. En yaygın kullanılan kayıp fonksiyonlarından biri olan ikili çapraz entropi (binary cross-entropy), çapraz entropi (cross-entropy) fonksiyonunun bir türüdür ve modelin ikili sınıflandırma problemlerindeki performansını ölçer. y , gerçek çıkış değerini ve $(\hat{y})$ tahmin edilen çıkış değerini temsil etmek üzere çapraz entropi (denklem 2.7) ve ikili çapraz entropi kayıp fonksiyonu (L) (denklem 2.8) şu şekilde hesaplanabilir:

$$L(y, \hat{y}) = - \sum_{c=1}^C y \log \hat{y} \quad (2.7)$$

$$L(y, \hat{y}) = -(y \log \hat{y} + (1 - y) \log(1 - \hat{y})) \quad (2.8)$$

Çapraz entropi, ağın tahminlerine göre eğitim setinin negatif logaritmik olasılığı olarak da tanımlanabilir. Başka bir deyişle, bu kayıp fonksiyonunu en aza indirmek, olasılığı en üst düzeye çıkarmaya karşılık gelir [94].

Ortalama mutlak hata (Mean Absolute Error, MAE), $\hat{y}$ ve y arasındaki fark olarak tanımlanır ve hataların mutlak ortalamasını ölçer. Dezavantajlarından biri aykırı değerlere ortalama karesel hata (Mean Squared Error, MSE) kadar duyarlı olmamasıdır. MAE ölçeğe bağlı olduğundan, farklı veri serileri üzerinde çalışan modelleri karşılaştırmak için kullanılması önerilmez [95]. Özellikle regresyon problemleri için kullanılan MSE ise hataların karelerinin ortalamasını ölçen en temel maliyet fonksiyonlarından biridir, temel olarak tahmin edilen değerler ile gerçek değerler arasındaki ortalama karesel farkı hesaplar. y , gerçek çıkış değerini ve $\hat{y}$ tahmin edilen çıkış değerini temsil etmek üzere MAE (denklem 2.9) ve MSE kayıp fonksiyonu (denklem 2.10) şu şekilde hesaplanabilir [95]:

$$MAE = \frac{1}{n} \sum_{i=0}^{i=n} |\hat{y}_i - y_i| \quad (2.9)$$

$$MSE = \frac{1}{n} \sum_{i=0}^{i=n} (\hat{y}_i - y_i)^2 \quad (2.10)$$

2.2.3. Geri Yayılım

Sinir ağlarının eğitimi geri yayılımla gerçekleştirilir. Geri yayılım, kayıp fonksiyonunun eğimini ağın parametrelerine göre verimli bir şekilde hesaplar. Rumelhart vd. [96] ilk kez 1986 yılında yayınlanan kitaplarında, geri yayılımla ilgili öncü başarılı deneylerin sonuçlarını sunmuştur ve bu sayede geri yayılımın popülerleşmesine büyük katkıda bulunarak, çok katmanlı sinir ağlarında aktif bir araştırma dönemi başlatmıştır [91]. İleri beslemeli sinir ağı kullanarak x 'i girdi kabul edip bir $\hat{y}$ çıkışı ürettiğimizde, bilgi

ağ boyunca ileri doğru akar. Girdi x , daha sonra her katmandaki gizli birimlere kadar yayılan ve $\hat{y}$ 'yi üreten ilk bilgiyi sağlar. Buna ileri yayılım denir. İleri yayılım, eğitim sırasında skaler bir maliyet (J) üretene kadar devam edebilir. Geri yayılım algoritması daha sonra eğimi hesaplamak için, maliyetten gelen bilginin ağ boyunca geriye doğru akmasına izin verir. Eğitim için analitik bir ifadenin hesaplanması kolay olsa da, böyle bir ifadeyi sayısal olarak değerlendirmek hesaplama açısından külfetli olabilir. Geri yayılım algoritması bu hesabı basit ve uygun bir prosedür kullanarak yapar. Geri yayılım teriminin, genellikle çok katmanlı sinir ağları için tüm öğrenme algoritması anlamına geldiği şeklinde yanlış bir yargı söz konusudur. Aslında, geri yayılım yalnızca eğimi hesaplama yöntemine atıfta bulunurken, gradyan iniş algoritması gibi başka bir algoritma bu eğimi kullanarak öğrenmeyi gerçekleştirmek için kullanılır [91]. Bir sinir ağı, bir sonraki tahmini iyileştirebilmek adına, geri yayılım ile mevcut tahminlerin doğruluğuna göre eğitilebilir parametreler üzerinde güncelleme yapar ve kaybı en aza indirecek parametre değerlerini (W, b) bulmayı amaçlar. Bu süreci optimize etmek amacıyla, maliyet fonksiyonu ($J(W, b)$) olarak adlandırılan, kayıp fonksiyonlarının ortalaması hesaplanır. n , veri kümesindeki örneklerin sayısı olmak üzere maliyet fonksiyonu şu şekilde hesaplanır:

$$J(W, b) = \frac{1}{n} \sum_{n=1}^n L(y, \hat{y}) \quad (2.11)$$

Gradyan iniş algoritması ile ağırlık (W) ve bias (b) parametrelerinin güncellenmesine dair matematiksel formüller denklem 2.12 ve denklem 2.13'te verilmiştir. Bu eğitilebilir parametrelere ilişkin maliyet fonksiyonunun türevi, ağırlık ve bias değerlerinin ne kadar güncellenmesi gerektiğini tahmin eder. Parametrelerin güncelleme oranını kontrol etmek için, öğrenme oranı (learning rate, α) adı verilen bir metrik kullanılır. Bu sayede değişim kontrol edilebilir. Yakın tahminlerde kayıp değerleri düşük olacaktır ve parametreler çok daha az güncellenecektir ancak aksi durumda yüksek kayıp değerleri için büyük değişiklikler gereklidir.

$$W := W - \alpha \frac{dJ(W, b)}{dW} \quad (2.12)$$

$$b := b - \alpha \frac{dJ(W, b)}{db} \quad (2.13)$$

2.2.4. Optimizasyon Algoritmaları

DL uygulamalarında, öğrenme işleminin sağlıklı bir şekilde sonuçlanması için, hata fonksiyonunun mutlak minimum değerinin bulunması gerekmektedir. Bu işlem optimizasyon yöntemleri kullanılarak gerçekleştirilir. Optimizasyon, ağıın ürettiği çıkış değeri ile gerçek değer arasındaki farkı, yani hatayı, minimize etmek için kullanılan yöntemlerdir. YSA'ların optimizasyonu için en çok kullanılan algoritmalarından biri, gradyan iniş algoritmasıdır. Tek iterasyonda kullanılan veri büyüklüğüne bağlı olarak üç adet gradyan iniş yöntemi (Batch Gradient Descent, Stokastik gradyan iniş (Stochastic Gradient Descent, SGD), Mini-Batch Gradient Descent) bulunur. SGD, gradyan iniş ile en düşük optimum noktayı bulana kadar bir fonksiyonu takip eden yinelemeli bir algoritmadır, ağıın tüm parametrelerini işlemek yerine rastgele örnekleri kullanarak parametreleri günceller ve bu sayede bellek verimliliği sağlanmasına yardımcı olur. İşlem, hesaplama açısından ucuzdur, ancak parametreler daha yüksek varyansa ve daha büyük dalgalanmalara sahiptir.

Gradyan iniş yöntemini esas alan RMSProp, Adagrad, Adam ve AdaDelta gibi birçok algoritma mevcuttur [92]. DSA'ların eğitiminde kullanılan en popüler yöntemlerden biri olan Adam algoritması, öğrenme oranı özel olarak belirlenebilen bir optimizasyon yöntemidir. İlk olarak 2014 yılında Kingma ve Ba [97] tarafından literatüre katılmış, SGD veya RMSProp'a kıyasla yüksek performansı nedeniyle çok sayıda araştırmacıdan yüksek ilgi görmüştür. Adam, her parametre için farklı öğrenme oranı belirleyebilmek için adaptif öğrenme hızı tekniklerini kullanır. Çok sayıda değişken veya kayıt içeren karmaşık problemlerle uğraşırken son derece verimlidir ve daha az belleğe ihtiyaç duyar. Momentumlu gradyan inişi (Gradient descent with momentum) ve RMSProp yöntemlerinin bir birleşimidir. Momentum yöntemi, gradyanların üssel ağırlıklı ortalamasını hesaba katarak gradyan iniş algoritmasını hızlandırır. Seyrek eğimli ortamlarda iyi performans göstermek için Adagrad [98]'in avantajlarını kullanır ve sinir ağlarının dışbükey olmayan optimizasyonu ile mücadele eder. Ayrıca Adagrad'ın eksikliklerini gidermek ve çevrimiçi ortamlarda iyi performans göstermek için RMSprop

[99] algoritmasının avantajını kullanır. AdaGrad'ın yaptığı gibi kümülatif kare gradyan sayısını hesaplamak yerine, bir 'üssel hareketli ortalama' hesaplar. RMSprop ve AdaGrad tekniklerinin güçlü özellikleri, daha optimize bir gradyan inişi sağlamak için Adam algoritması tarafından miras alınır. AdaMax ise yine ilk kez Adam ile aynı çalışmada [97]], Adam algoritmasının bir varyasyonu olarak karşımıza çıkar. Adam algoritması, ağırlıkları güncellemek için L_2 normu kullanırken, AdaMax sonsuz normu (L_∞) kullanır. Sonsuz norm, gradyan büyüklüğünün daha sağlam bir ölçüsünü sağlar ve bu sayede gradyanların kaybolması veya sonsuza gitmesi gibi sorunları hafifletmeye yardımcı olur. Ayrıca aşırı gradyanların daha iyi işlenmesini sağlaması sebebiyle, özellikle karmaşık mimarilere sahip DL modellerinde eğitim sürecini stabilize etmeye yardımcı olur.

2.3. Derin Öğrenme Modelleri

2.3.1. Otokodlayıcı (Autoencoder)

İlk olarak Rumelhart vd. [96] tarafından, girdisini yeniden oluşturmak üzere eğitilen bir sinir ağı olarak tanımlanan Otokodlayıcılar, genellikle giriş ve çıkış verileri eş sayıda olan, boyut indirgeme ve öznetelik çıkarımı için kullanılan özel bir sinir ağı çeşididir [100]. Otokodlayıcılar eğitim sırasında veri etiketlerine ihtiyaç duymadığından, denetimsiz öğrenme için kullanılan ileri beslemeli bir sinir ağı olarak kategorize edilir. Bir otokodlayıcı, temel olarak aşağıdaki görevleri gerçekleştiren bir kodlayıcı (encoder) ve bir kod çözücü (decoder) olmak üzere iki aşadan oluşur:

- Kodlayıcı: Yüksek boyutlu giriş verilerinin temel özelliklerini öğrenir ve girdiden daha düşük boyutlara sahip olan gizli uzay katmanında sıkıştırır.
- Kod Çözücü: Gizli uzaydan gelen bilgilerle giriş verilerini yeniden oluşturmaya çalışır.

Matematiksel olarak $f(\cdot)$ kodlayıcı fonksiyonu, $g(\cdot)$ kod çözücü fonksiyonu ve X giriş vektörü olmak üzere, çıkış ($\hat{X}$) basitçe aşağıdaki gibi tanımlanabilir:

$$g[f(X)] = \hat{X} \quad (2.14)$$

Otokodlayıcılar geri yayılıma dayalı denetimsiz bir öğrenme algoritması kullanılarak eğitilir. Kodlayıcı, girdi değerlerini bir gizli uzay temsiline (kod (h)) sıkıştırırken, kod çözücü, girdiyi gizli uzaydan yeniden oluşturmaya çalışır. Bu işlemi yinelemeli olarak çalıştıran model, çıkış değerlerini mümkün olduğunca girdiye benzer tutacak bazı optimal parametreleri (ağırlıkları) bulacaktır. Kodlama işlemi, denklem 2.15'te verilmiştir [101]. Burada W_1 giriş katmanı ile gizli katman arasındaki ağırlık matrisi, b_1 bias vektörü, $f(x)$ ise doğrusal olmayan aktivasyon fonksiyonudur.

$$h = f(x) = f(W_1X + b_1) \quad (2.15)$$

Bu yaklaşımın işe yaramasının sebebi, çoğunlukla gizli uzay vektörünün orijinal girdiden çok daha küçük olmasıdır. Böylece, önemli bilgiler kodlanırken gürültü otomatik olarak kaldırılır. Kod çözme işlemi, denklem 2.16'da verilmiştir. Burada, W_2 gizli katman ve çıkış katmanı arasındaki ağırlık matrisi, b_2 bias vektörü ve $g(x)$ doğrusal olmayan aktivasyon fonksiyonudur.

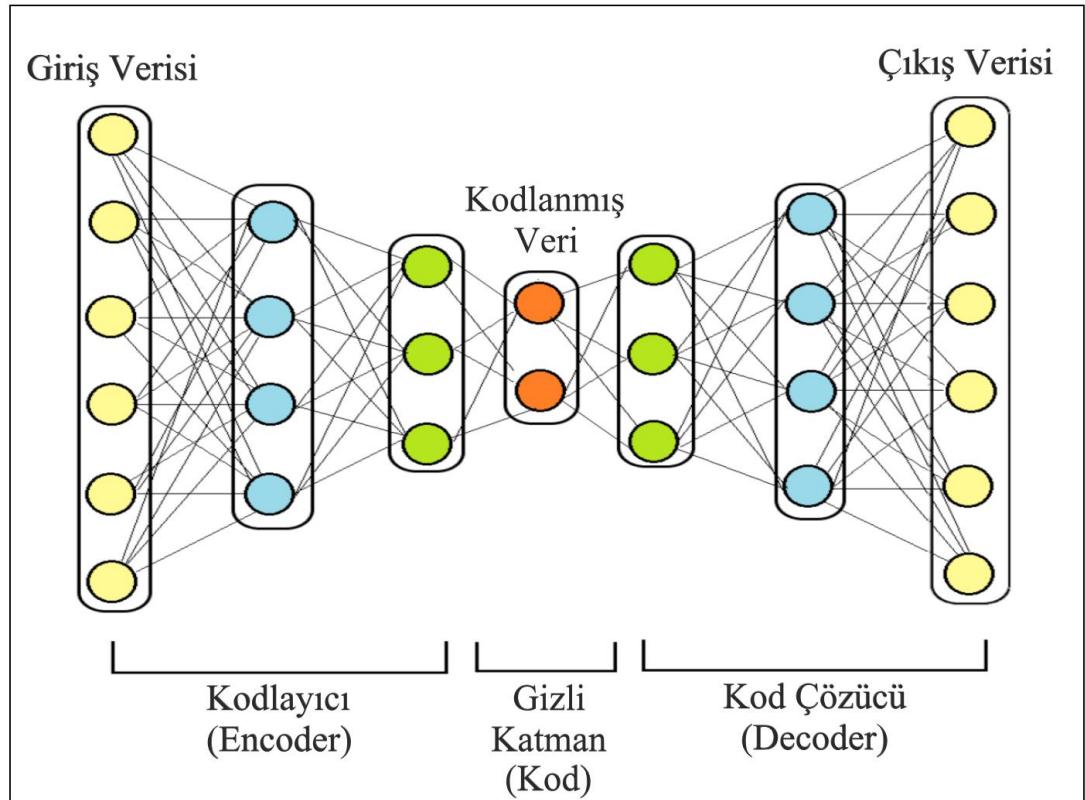
$$\hat{X} = g(x) = f(W_2h + b_2) \quad (2.16)$$

Kod çözme işleminden sonra, çıktı vektörü girdi vektörüne oldukça benzer bir şekilde (neredeyse bozulma olmadan) temsil edilebilir. Modelin temel eğitim amaçlarından ilki, X ve $\hat{X}$ arasındaki yeniden yapılandırma hatasını (L) en aza indirmektir [101]. Giriş ve çıkış arasındaki kayıp fonksiyonunu (J) ölçmek için çapraz entropi ve ortalama kare hata gibi çeşitli metrikler kullanılır. Diğer temel amaç, modelin giriş verilerini ne derece iyi yeniden oluşturabildiğini gösteren kaybı azaltmak (kayıp fonksiyonu) ve gürültüyü filtrelerken verilerin en önemli özelliklerini yakalamaktır. Otokodlayıcının yeniden yapılandırma hatası ve ortalama karesel hata kullanılarak kayıp fonksiyonu hesaplaması, giriş verileri: $X = x_1, x_2, \dots, x_n$ ve çıkış verileri: $\hat{X} = \hat{x}_1, \hat{x}_2, \dots, \hat{x}_n$ olmak üzere, denklem 2.17 ve denklem 2.18'te verilmiştir.

$$L(X, \hat{X}) = \|\hat{X} - X\| \quad (2.17)$$

$$J(X, \hat{X}) = \frac{1}{2} \sum_{i=1}^n \|\hat{X}_i - X_i\|^2 \quad (2.18)$$

Otokodlayıcılar, günümüzde birçok uygulama için kullanılabilen popüler sinir ağı ailesidir. Gelişmiş doğrusal olmayan veri filtreleme, gürültü giderme, boyut azaltma ve veri mühendisliği vb. için yaygın olarak kullanılırlar [102]. Otokodlayıcıların pek çok farklı modifikasyonu vardır, en temel otokodlayıcı mimarisi Şekil 2.3'te gösterilmiştir.



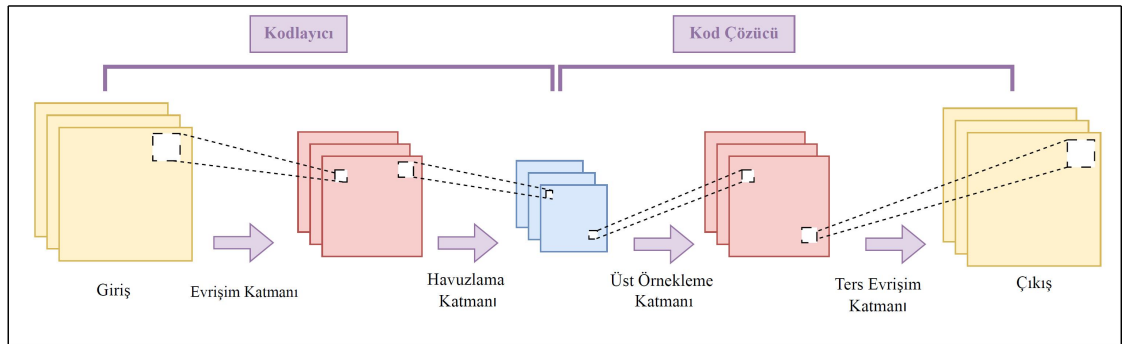
Şekil 2.3. Temel bir otokodlayıcı mimarisi.

Otokodlayıcı ve PCA yöntemleri, boyut azaltma açısından benzer etkilere sahip olmasına rağmen otokodlayıcı PCA'dan daha esnektir [101]. Otokodlayıcı sinir ağlarının temel avantajları; geniş verilere uyarlanabilir olmaları, uygulanması basit, eğitilmesi kolay ve hesaplama açısından verimli olmalarıdır. Otokodlayıcı tasarlarken en temel problemlerden biri, giriş verilerine olan duyarlılığıdır. Bir otokodlayıcı, temel veri özelliklerini mümkün olduğunca doğru bir şekilde yansıtan bir temsili öğrenirken, aynı zamanda orijinal eğitim setinin ötesinde genelleşen özellikleri kodlayabilmeli ve diğer veri setlerinde benzer özellikleri yakalayabilmelidir. Bu nedenle, otokodlayıcıların ilk

tanıtılmasından bu yana çeşitli varyantlar önerilmiştir. Bu varyantların temel olarak daha gelişmiş genelleme ve giriş verilerine olan bağımlılığı azaltma gibi farklı amaçları vardır. Bazı önemli örnekler arasında Gürültü Giderici Otokodlayıcı [103], Seyrek Otokodlayıcı [104], Varyasyonel Otokodlayıcı [105] ve Evrişimli Otokodlayıcı [106] bulunmaktadır.

2.3.1.1. Evrişimli Otokodlayıcı

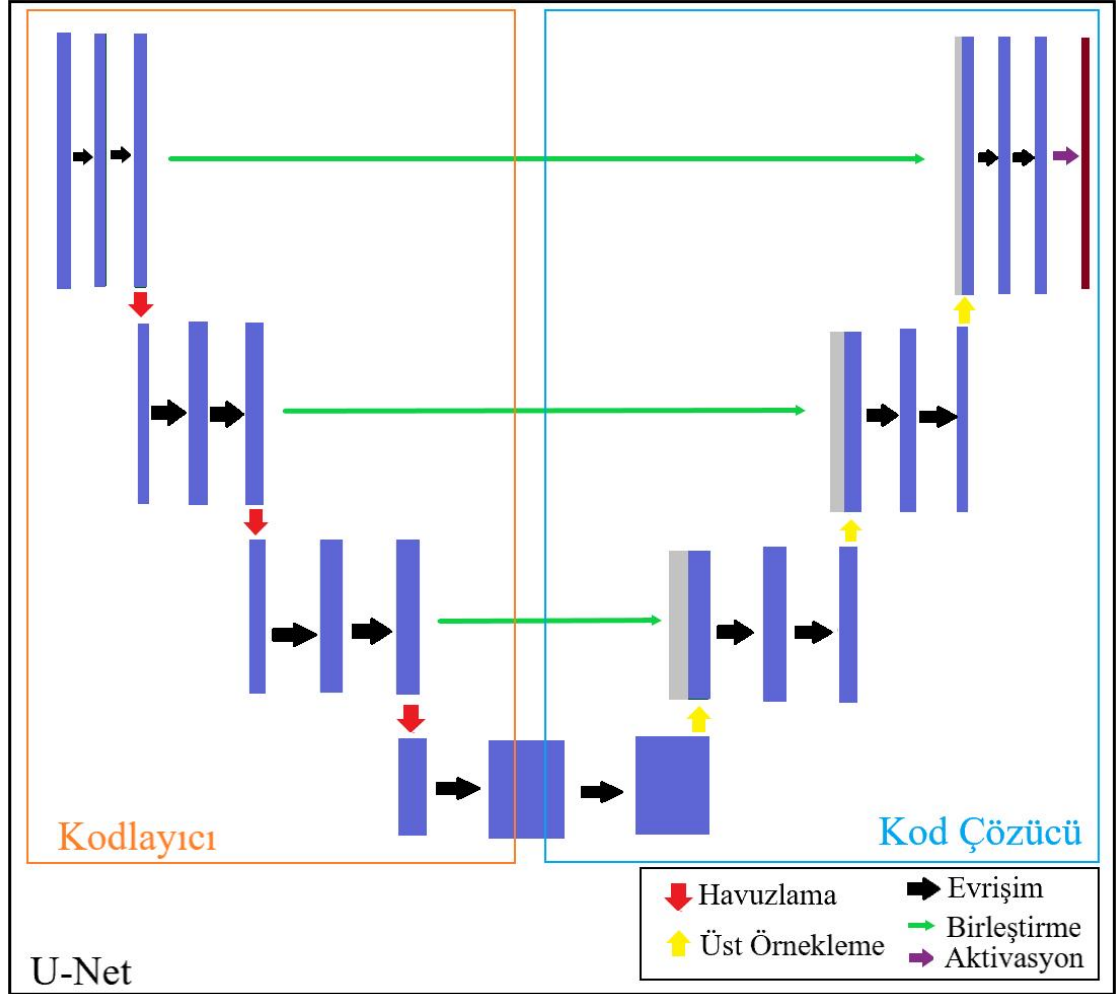
Geleneksel otokodlayıcılar genellikle tam bağlı YSA'lar ile oluşturulur, fakat görüntü verileri gibi çok boyutlu veriler için giriş verilerinin işlenmeden önce tek boyutlu vektörlere dönüştürülmesi gerekir, bu da görüntünün iki boyutlu yapısının değişmesi sebebiyle bilgi kaybına neden olur. Evrişimli otokodlayıcılar (Convolutional Autoencoder, CAE), geleneksel otokodlayıcılardaki tam bağlı katmanların yerini CNN'deki evrişim (convolution) ve havuzlama (pooling) işlemleri alacak şekilde tasarlanmıştır. Böylece CAE, iki boyutlu bilginin kaybını önleyerek daha iyi öznetelik çıkarımı yapar [101]. CAE mimarisinde de otokodlayıcı ile aynı şekilde çıkış katmanının boyutu giriş katmanı ile aynıdır, ancak kodlayıcı ağ evrişim katmanlarına ve kod çözücü ağ da ters evrişim (deconvolution) katmanlarına dönüşür. CAE'lerin eğitimi, giriş verilerinin evrişim kullanılarak kodlanması ve daha sonra bu kodun kod çözücü ağı ile çözülmesi sürecini kapsar. Giriş sinyalinin özelliklerini çıkarmak için evrişim ve havuzlama kullanılır ve ağırlıklar paylaşılır. Kod çözücü ağı, düşük boyutlu özelliklerin, veriyle aynı boyutta yeniden yapılandırılması için ters evrişim ve üst örnekleme (upsampling) katmanlarını içerir. Temel bir CAE mimarisi şekil 2.4'te gösterilmiştir.



Şekil 2.4. Temel Evrişimli Otokodlayıcı (CAE) mimarisi.

2.3.2. U-Net

U-Net, ilk kez Ronneberger vd. [107] tarafından 2015 yılında, biyomedikal görüntü segmentasyonu için geliştirilen bir CNN mimarisidir. U-Net modeli adını, CNN katmanları tarafından oluşturulan ayırt edici "U" şeklinden alır. Temelde, sol kısımda daralan alt örnekleme yolu ve sağ kısımda genişleyen üst örnekleme yolu olmak üzere iki ana bölümden oluşur. İlk bölümde, otokodlayıcıdaki kodlayıcı ağına



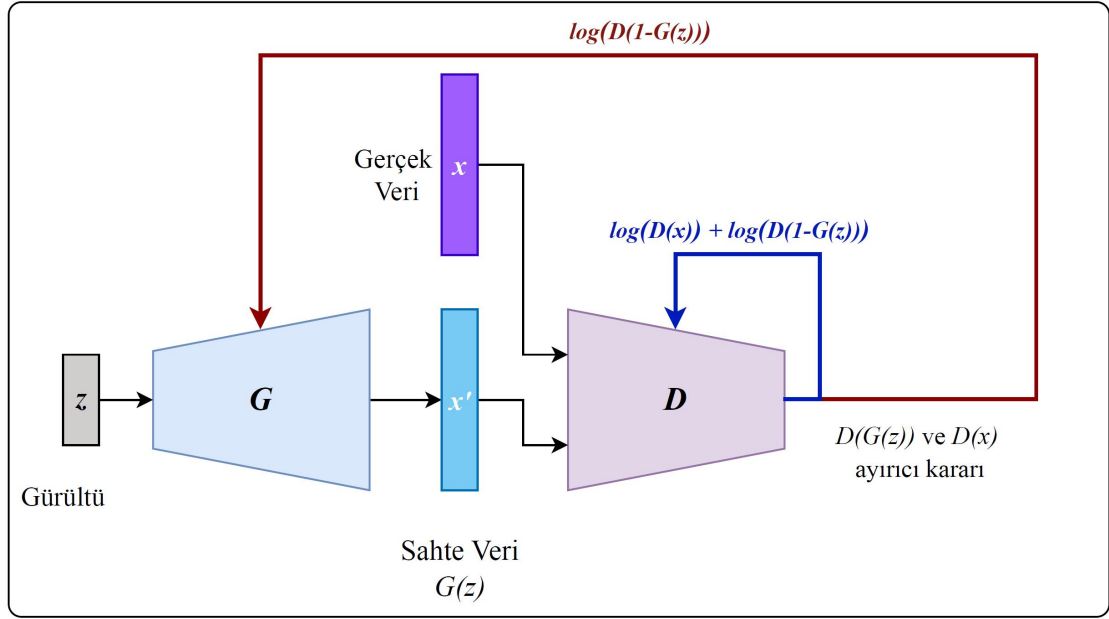
Şekil 2.5. Temel U-Net mimarisi.

benzer şekilde, havuzlama katmanları ile boyut azaltma uygulanır ve CNN'ler ve aktivasyon fonksiyonları ile giriş verisinden gelen bilgiler öğrenilir. İkinci bölümde ise, bir kod çözücü gibi hareket ederek boyutu artıran teknikler kullanır ve çıktının kalitesini iyileştirir. Tüm bunları yaparken, önceki aşamalardan gelen önemli bilgileri, atlama bağlantılarıyla kod çözücü aşamalarına iletir. Ağın merkezinde, giriş verisinin temel kısmını yakalayan en sıkıştırılmış öznitelik gösterimi elde edilir. Şekil 2.5, temel bir U-Net mimarisini göstermektedir. U-Net ağları, literatürde sıklıkla görüntü

segmentasyonu için kullanılmaktadır [108]. Tek/çok yönlü ağların sınırlamalarını ele almak için birçok model, görüntüden segmentasyon haritasına uçtan uca eğitim kuran kodlayıcı-kod çözücü yollarına sahip U-Net varyantlarını kullanır [109]. Kodlayıcı, tek/çok yönlü ağlara benzerdir ancak öznetelik haritalarının farklı ölçekleri arasında aşağı örnekleme işlemleri vardır. Kod çözücü, kodlayıcıdan çıkarılan öznetelikleri kullanır ve orijinal görüntüyle aynı boyutta bir çıktı üretir. Öznetelik haritalarını kodlayıcıdan doğrudan kod çözücüye geçiren atlama bağlantıları U-Net'in performansına katkıda bulunur. Aktarılan bilgiler, ayrıntıların kurtarılmasına yardımcı olabilir. U-Net'in en yaygın modifikasyonları, artık bloklar [110], yoğun bloklar [111], dikkat modülleri [112] vb. gibi diğer evrişimli modülleri içerir. Bu evrişimli modüller, düzenli evrişim işlemlerinin yerini alabilir veya U-Net'in atlama bağlantılarında kullanılabilir. Artık bloklar, modülün girişini çıkışına ekleyerek eğitim sırasında gradyanların kaybolması sorununu hafifletir ve bu da yakınsama hızına katkıda bulunur [110].

2.3.3. Üretici Çekişmeli Ağlar

Üretici modeller, bir veri sınıfına ait dağılımı oldukça efektif bir şekilde içsel özneteliklerin fonksiyonu olarak ifade edebilmektedir. Dolayısıyla normal ve anormal verileri yüksek doğrulukla ayırt edebilme, farklı alanlara ait verilerin ortaklıklarını bulup birbirine dönüştürebilme gibi, daha önce DL alanında gerçekleştirilememiş buluşların ortaya çıkmasına önayak olmuştur. Üretici ağlar, CNN'lere kıyasla, amaç ve yöntemleri bakımından temelde farklıdır. Başlangıçta bir görüntü analiz aracı olmaları değil, doğal görünümlü görüntüler üretmeleri amaçlanmıştır [113]. Daha sonra, üretken yöntemleri dengelemek için çekişmeli eğitim paradigması önerilmiştir. Üretici Çekişmeli Ağlar (Generative Adversarial Networks, GAN) ilk olarak 2014 yılında Goodfellow vd. [114] tarafından üretici modellerin kısıtlamalarını ortadan kaldırmak amacıyla tanıtılmıştır. GAN'lar, bir ağ parçasının bir göreve çözümler ürettiği ve diğer bir parçanın üretilen çözümleri, önceden bilinen çözümlerle karşılaştırdığı ve derecelendirdiği bir tür sinir ağı mimarisidir. Çekişmeli ağda üretici modele (generator, G), bir örneğin model dağılımından mı yoksa veri dağılımından mı olduğunu belirlemeyi öğrenen rakip bir ağ (ayırıcı model (discriminator, D)) eşlik eder. Temel GAN mimarisinin temsili bir görseli şekil 2.6'da verilmiştir.



Şekil 2.6. Temel Üretici Çekişmeli Ağ (GAN) mimarisi.

GAN, üreticinin giderek daha gerçekçi veriler üreterek, ayırıcı ağın üretilen verileri sahte olarak tanımlama olasılığını en aza indirmeye çalıştığı, ayırıcının ise gerçek ve üretilen veriler arasında ayırım yapma yeteneğini geliştirerek bu olasılığı en üst düzeye çıkarmaya çalıştığı çekişmeli bir kayıp kullanılarak eğitilir. Eğitim süreci, her iki ağın da gerçekçi veriler üretilene kadar performanslarını iyileştirdiği yinelemeli bir oyundur. Bu oyundaki rekabet, her iki ağı da sahte ürünler gerçek ürünlerden ayırt edilemez hale gelene kadar yöntemlerini geliştirmeye yönlendirir ve bu da üretilen çözümleri gerçek çözümlerden ayırt etmek üzere eğitilebileceği ve görevinde başarılı olduğu sürece, üretici için bir eğitim sinyali türetilebileceği anlamına gelir. Çekişmeli/düşman kavramı GAN ismine böyle girmiştir. Ayırıcı ağ, gerçek çözümleri üretilen çözümlerden ayırt etmek üzere eğitilirken, üretici parça en gerçekçi görünen çözümlere ulaşmak üzere eğitilir ve bu da onları amaçları açısından düşman yapar. GAN'ın eğitim algoritması aşağıdaki gibidir [115]:

1. Ayırıcı ağın eğitimi:

- Eğitim veri kümesinden rastgele bir x örneği alınır.
- Yeni z rasgele gürültü vektörü alınır ve üretici ağ kullanılarak $\hat{x}$ sahte örneği sentezlemek için kullanılır.

- x ve $\hat{x}$ 'i sınıflandırmak için ayırıcı ağ kullanılır.
- Sınıflandırma hatalarını hesaplayıp minimize etmek ve ayırıcı ağırlıklarını güncellemek için toplam hata geriye doğru hesaplanır.

2. Üretici ağı eğitimi:

- Yeni z rasgele gürültü vektörü alınır ve üretici ağ kullanılarak $\hat{x}$ sahte örneği sentezlenir.
- $\hat{x}$ 'i sınıflandırmak için ayırıcı ağ kullanılır.
- Sınıflandırma hatasını bulmak, ayırıcı hatasını maksimize etmek ve üretici ağırlıklarını güncellemek için hata geriye doğru hesaplanır.

Eğitim sonunda GAN tarafından üretilen sahte verilerin dağılımı (p_{sahte}) ile gerçek verilerin dağılımının (p_{gercek}) birbirine çok yakın olması beklenir. Buna göre, GAN sahip olduğu iki ayrı ağı minimize ve maksimize etmek için aşağıdaki hesaplamayı kullanır:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{gercek}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log 1 - D(G(z))] \quad (2.19)$$

Denklem 2.19'da, ilk toplam ifadesi, D ayırıcı ağı gerçek görüntülerin (x) orijinal olduğunu tahmin etme olasılığını hesaplarken, ikinci toplam ifadesi, z gürültüsü ile D 'nin G üretici ağı ürettiği görüntülerin sahte olduğunu tahmin etme olasılığını göstermektedir. GAN modelleri, bir alana ait örneklerden oluşan veri özelliklerini mümkün olan en üst hiyerarşik seviyede öğrenerek aynı alana ait istenilen özelliklerde sentetik veri üretebilmektedir. Bunu gerçekleştirmek için, söz konusu veri sınıfının içsel yapısını ve öznitelik dağılımını en iyi şekilde keşfetmeye çalışır.

GAN, önerilen temel çalışmasından bu yana DL çalışmalarının önemli bir kolu olmuştur. GAN tabanlı modeller, diğer üretici modellerle (ör: varyasyonel otokodlayıcı) karşılaştırıldığında üstün performansları sayesinde, görüntü sentezi (örneğin, İnsan yüz görüntüleri [116]), metin oluşturma (örneğin, görsel paragraf oluşturma [117]) ve müzik sentezi (örneğin, müzik kompozisyonu [118]) gibi geniş bir uygulama alanına sahiptir. Son zamanlarda, GAN modelleri biyomedikal araştırmalar için de uyarlanmış, örneğin dizi verilerinin (T-hücresi reseptör dizileri [119]) oluşturulmasında ve tıbbi görüntülemenin geliştirilmesinde [120] kullanılmıştır. Ayrıca, yüksek seviyeli öznitelikler

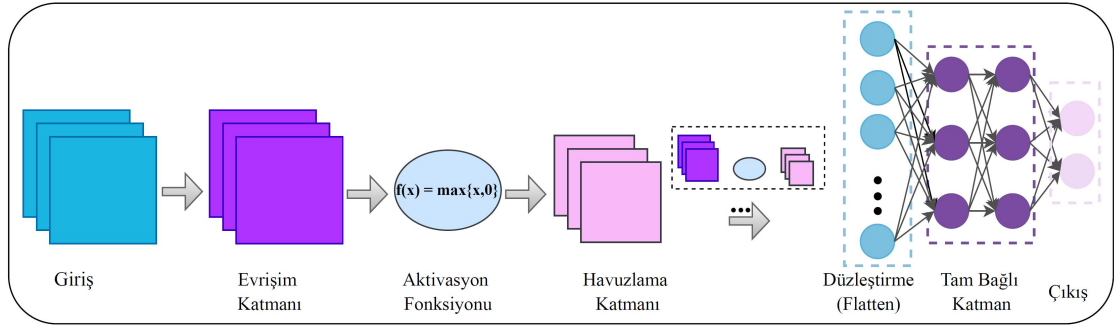
elde ederek veri yapısı ve dağılımını modelleyebilmesi ve buna ek olarak çekişmeli eğitim aracılığıyla öğrenilen veri dağılımından örnekleme yapılmasına olanak tanınması nedeniyle araştırmacıların dikkatini çekerek çok sayıda GAN tabanlı modelin ortaya çıkmasına neden olmuştur. Yarı Denetimli GAN [115], Derin Evrişimli GAN [116], Aşamalı GAN [120] ve Koşullu GAN [115] gibi farklı amaçlar için geliştirilmiş birçok GAN tabanlı yöntem bulunmaktadır.

2.4. Sınıflandırıcı Mimarileri

2.4.1. Evrişimli Sinir Ağları

Evrişimli sinir ağları (CNN'ler), giriş verilerini işlemek için evrişim işlemi kullanan özel bir derin sinir ağıdır. 1980'li yıllarda, büyük ölçüde nörobilimden ve görsel korteksin organizasyonundan ilham alınarak ortaya çıkarılmış, 1989 yılında LeCun vd. [121] tarafından yapılan çalışmadan günümüze kadar birçok alanda yaygın olarak kullanılmıştır. CNN'ler derin, tam bağlantılı ağların aksine, ağırlık paylaşım özellikleri sayesinde eğitilebilir parametrelerin sayısını azaltırken, aynı zamanda uzamsal veya ızgara benzeri bir organizasyona sahip verileri (örneğin, zaman serileri, görüntüler, videolar, vb.) işlemede oldukça başarılıdır [94]. Temel CNN mimarisi genellikle iki bölümden oluşur: ilki evrişim işlemlerine dayanır ve görüntü için özellikleri öğrenir, ikinci bölüm ise özellikleri düzleştirir ve bunları sınıflandırma/regresyon işlemlerini gerçekleştirmek için tam bağlı katmanlar kümesine (başka bir deyişle, çok katmanlı bir algılayıcıya) aktarır. Bu iki bölüm ardışık olarak uçtan uca eğitilir. İlk bölümde, bir evrişim işlemi, doğrusal olmayan bir aktivasyon fonksiyonu (örneğin, bir ReLU) ve bir havuzlama işleminden oluşan bir dizi bloğu birleştirilir. Giriş ve çıkış katmanı, evrişimli katmanlar, havuzlama katmanı ve tam bağlantılı katmanlardan oluşan temel bir CNN mimarisi örneği şekil 2.7'de verilmiştir.

Evrişim işleminin temel amacı, giriş verilerinden öznitelik çıkarımını gerçekleştirmektir. Bunun için filtreler olarak da adlandırılan çekirdeklerden (kernel) yararlanır. Matematiksel olarak evrişim işlemi, girişin filtre boyutundaki bir bölümü ile filtre vektörünün noktasal çarpım sonuçlarının toplanmasını içerir. Çoğunlukla giriş verilerinin boyutu, filtre boyutundan (3x3 gibi) çok daha büyüktür [97]. Evrişim işlemi, giriş



Şekil 2.7. Temel CNN mimarisi.

matrisinin tüm elemanları gezilene kadar, filtre konumu belli bir adım sayısı (stride) kadar değiştirilerek tekrarlanır. Bu şekilde her parça için bir değer hesaplanır ve yeni bir çıkış matrisi oluşturulur. Evrişim işlemi giriş verisinin boyutunu azaltabilir, ters evrişim ise düşük boyutlu bir girdi alır ve daha yüksek boyutlu yeni bir matris oluşturur, bu sebeple üst örnekleme yapmak için ve otokodlayıcı, U-Net gibi modellerin kod çözücü ağında kullanılır.

Havuzlama işleminin temel amacı, anlamsal olarak benzer öznitelikleri bir araya getirmektir. Bunun için giriş verisi, evrişim ve aktivasyon işlemlerinden geçtikten sonra havuzlama katmanına gelir ve burada havuzlama işlemi ile temsili öznitelikler korunurken alt örnekleme öznitelik haritaları oluşturulur ve giriş verisinden çıkarılan öznitelikler konumlara göre öznitelik haritalarına kaydedildikten sonra boyut azaltma ve öznitelik iyileştirme sağlanır [122]. Havuzlama işlemi temel olarak iki farklı şekilde yapılabilir. Normal havuzlama işlemi ile girdinin bir kısmına havuzlama uygulanırken, küresel havuzlama işlemi ile girdinin tüm değerlerine aynı anda havuzlama yapılır. Bu temel kategorilendirmenin altında farklı amaçlarla kullanılan üç farklı havuzlama çeşidi bulunur. Maksimum havuzlamada daha yüksek, minimum havuzlamada ise daha düşük değerler değiştirilmeden iletilir. Ortalama havuzlamada ise giriş değerlerinin ortalama değeri alınarak çıkış değeri elde edilir [97].

CNN'ler öncelikle, çoklu diziler biçiminde gelen, örneğin üç renk kanalında piksel yoğunlukları içeren 2 boyutlu (2B) renkli bir görüntü gibi verileri işlemek için tasarlanmıştır. Birçok veri biçimi çoklu diziler biçimindedir, genellikle dil ve sinyal işleme dizileri için 1B, görüntüler veya ses spektrogramları için 2B ve video veya

hacimsel görüntüler için 3B-CNN'ler kullanılır [122].

2.5. Transfer Öğrenme

Transfer öğrenme üzerine yapılan araştırmalar, 1995'ten [123] beri; öğrenmeyi öğrenme, bilgi transferi, tümevarımsal transfer, çoklu görev öğrenmesi, bağlam duyarlı öğrenme ve bilgi tabanlı tümevarımsal önyargı gibi farklı isimler altında, giderek daha fazla ilgi görmektedir [124]. Transfer öğrenme, önceki görevlerden gelen bilgiyi, az miktarda yüksek kaliteli eğitim verisine sahip hedef göreve aktarmaya çalışır ve yüksek boyutlu, düşük örneklem büyüklüğüne sahip verilerin getirdiği dezavantajların üstesinden gelmeyi amaçlar [125]. Transfer öğreniminde, bir eğitim ortamında bir model tarafından öğrenilen alan bilgisi kullanılır ve bu bilgi, farklı ama ilgili bir probleme uygulanır. Bu yaklaşım, model belirli bir alandan büyük miktarda veri (örneğin, biyolojik diziler) ile eğitildiğinde, göreve özel özelliklere ek olarak alan için genel özellikleri de öğreneceğinden oldukça etkilidir. Bu önceden eğitilmiş model, daha sonra tekrar eğitilebilir veya ince ayarlanabilir, böylece çözülmek istenen bir alt akış görevini, daha az göreve özel veri kullanarak, daha kısa eğitim süresiyle ve daha doğru bir şekilde tamamlayabilir [125]. Önceden eğitilmiş modeller, genellikle az miktarda veya nispeten karmaşık verileri modellemek için, sınırlı hesaplama kaynaklarına sahip olan akademik ve özel sektör uygulamaları için kullanışlıdır. DL yaklaşımları, özellikle çeşitli ve yüksek boyutlu verilerde birden fazla özellik ve yardımcı değişken arasındaki karmaşık ilişki kalıplarını tespit edebilecek şekilde tasarlanmıştır. Ayrıca, DSA'nın genelleştirilebilir kalıpları öğrenme ve daha sonra bu genel modeli çalıştırma veya sisteme özgü görevlerde ince ayarlama süreci olan transfer öğrenimini gerçekleştirebildiği kanıtlamıştır [126].

David M. M. vd. [126], mikrobiyom verilerinde bulunan temel etkileşimleri daha iyi anlamak için, doğal dil işlemede (natural language processing, NLP) transfer öğrenmesi yeteneklerini kanıtlamış DL modellerinin kullanılabilirliğini öne sürer. Çünkü doğal dil verileri ile mikrobiyom verileri arasında kolayca görülebilen paralellikler bulunur, örneğin belgeler biyolojik örneklerle, kelimeler taksonlara ve konular mikrobiyal komşuluklara benzerdir. Bu konuda literatürde bulunan örneklerden birinde [127], basit bir NLP algoritması ile bireysel takson sayımlarına dayalı analiz yapılmış ve İBH ile ilişkili mikrobiyomlar ile sağlıklı bağırsak mikrobiyomları arasında doğru ve

genelleştirilebilir bir ayırım yapılabilirdiği gösterilmiştir. Bir başka yaklaşım da, modeli büyük ve çeşitli bir çerçevede özdenetimli öğrenme stratejileri kullanarak önceden eğitmektir. NLP’de bu yaklaşım, giriş metninden kelimeleri engelleyip, modelden gizli kısmı yeniden oluşturmalarını isteyerek veya modelden cümledeki bir sonraki kelimeyi üretmesini isteyerek uygulanmaktadır. Bu yaklaşım, modeli eksik bilgileri çıkarabilmek için giriş verilerindeki yararlı özellikleri belirlemeye zorlar ve modelin basit kelime birlikteliklerinin ötesinde, doğal metnin bileşimini yöneten dilbilimsel ve anlamsal kalıpları öğrenmesine yol açabilir (örneğin, dil bilgisi kuralları ve anlamsal ilişkiler). Benzer şekilde özdenetimli öğrenme, farklı mikrobiyal türler veya metabolik süreçler arasındaki etkileşimleri yakalayan ve mikrobiyal toplulukların paylaşılan kurallarını anlamak için kullanılabilen bir mikrobiyom modelini öğrenmek için uygulanabilir. Bu tür modeller, mikrobiyal sistemler arasında genelleme yapmak, girdi boyutluluğunu azaltmak ve mikrobiyal topluluklarda yararlı gizli özellikler bulmak için kullanılabilir [126].

2.6. Performans Değerlendirme Metrikleri

DL çalışmalarında kullanılan performans ölçütleri, söz konusu modelin sınıflandırma yeteneklerinin değerlendirilmesinde önemli bir rol oynar. Hata matrisi (confusion matrix) olarak tanımlanan değerlendirme protokolü, sınıf sayısına eş boyutlara sahip olan, sınıflandırma için kullanılan modelin tahmin sonuçlarını özetleyen bir kare matristir. Hata matrisi; gerçek pozitif (true positive, TP), gerçek negatif (true negative, TN), yanlış pozitif (false positive, FP) ve yanlış negatif (false negative, FN) bilgilerine ihtiyaç duyar. TP, gerçekte pozitif olan ve model tarafından da pozitif olarak doğru tahmin edilen örnekleri ifade ederken, FP, gerçekte negatif olan, ancak model tarafından pozitif tahmin edilen örnekleri gösterir. TN hem gerçekte negatif olan hem de model tarafında negatif olarak doğru tahmin edilen örneklerin sayısını gösterirken FN gerçekte pozitif olan, ancak model tarafından negatif olarak tahmin edilen örnekleri temsil eder. Hata matrisini kullanarak hesaplama yapan birçok değerlendirme metriği bulunmaktadır. Bu ölçütler temel olarak, eşik bağımlı metrikler ve eşikten bağımsız metrikler olmak üzere ikiye ayrılır. Eşik bağımlı metrikler, sınıflandırıcı performansını 0 ile 1 arasında belirli bir karar eşiğinde değerlendirir. Eşik ayarı, sınıflandırıcı kalitesini değiştirmez ancak performans metriğini iyileştirebilir, aynı zamanda sınıf dengesizliğini ele almanın en basit yollarından biridir

[10].

Doğruluk (accuracy): Doğru şekilde sınıflandırılan örneklerin, toplam örnek sayısına oranını hesaplar, ancak tüm örneklerin çoğunluk sınıfına ait olduğunu tahmin ederek yüksek bir doğruluk puanı elde edebilmesi sebebiyle sınıf dengesizliği altında yanıltıcı olabilir. Basitçe şu şekilde hesaplanır:

$$\text{doğruluk} = \frac{TP + TN}{TP + TN + FP + FN} * 100 \quad (2.20)$$

Duyarlılık (Recall): Bir veri kümesi için, tüm pozitif örnekler arasındaki gerçek pozitif tahminlerin oranını ölçer. Yanlış negatiflerin maliyetinin yüksek olduğu durumlarda yararlanılır. Bir modelin sınıflandırma duyarlılığı şu şekilde hesaplanabilir:

$$\text{duyarlılık} = \frac{TP}{TP + FN} * 100 \quad (2.21)$$

Kesinlik (Precision): Bütün pozitif tahminler arasında gerçek pozitif tahminlerin oranını ölçer. Denklem 2.22’de gösterildiği gibi, gerçek pozitiflerin sayısının, gerçek pozitiflerin ve yanlış pozitiflerin toplamına bölünmesiyle hesaplanır.

$$\text{kesinlik} = \frac{TP}{TP + FP} * 100 \quad (2.22)$$

Özgüllük (Specificity): Özgüllük, tüm negatif örnekler arasındaki gerçek negatif tahminlerin oranını ölçer. Spesifiklik/Özgüllük, olumsuz örneklerin doğru şekilde tanımlanmasının önemli olduğu uygulamalarda kullanılır. Özgüllüğe ilişkin matematiksel hesaplama aşağıdaki gibidir:

$$\text{Özgüllük} = \frac{TN}{TN + FP} * 100 \quad (2.23)$$

F1 Puanı: F1 puanı hem yanlış pozitifleri hem de yanlış negatifleri dikkate alan kesinlik ve duyarlılığın harmonik ortalaması olarak tanımlanır. Bu metrik özellikle sınıf dağılımı dengesiz olduğunda veya farklı hata türleriyle ilişkili değişken maliyetler söz konusu olduğunda faydalıdır. F1 puanına ilişkin matematiksel hesaplama aşağıdaki gibidir:

$$F1 \text{ Puanı} = \frac{2 * kesinlik * duyarlılık}{kesinlik + duyarlılık} * 100 \quad (2.24)$$

ROC Eğrisi Altındaki Alan (AUC): ROC (Receiver Operating Characteristic) eğrisi, çeşitli eşik ayarlarında, gerçek pozitif oran-yanlış pozitif oran değişim grafiğini çizer. AUC değeri ise ROC eğrisinin altındaki alanı ölçer ve modelin sınıflar arasında ayırım yapma yeteneğinin toplu bir ölçütünü sağlar. AUC, bir sınıflandırıcının tüm olası karar eşikleri boyunca pozitif ve negatif örnekler arasında ayırım yapma yeteneğini niceliksel olarak belirleyebilen, eşikten bağımsız bir metriktir. Başka bir deyişle, sınıflandırıcının ayırım yapma kapasitesini ölçer. AUC, sınıf dengesizliğine karşı dayanıklıdır ve yanlış pozitif ve yanlış negatiflerin göreceli maliyetleri eşit olduğunda kullanışlıdır [10].

Değerlendirme metrikleri genel olarak bir modelin sınıflandırma performansının kapsamlı bir analizini sağlamaktadır. Sınıflandırma görevlerinde sınıfların dağılımı eşit olmayabilir ve bir sınıfın sayısı diğerlerinden önemli ölçüde fazla olabilir. Uygun bir performans metriğinin seçilmesi, görevin özel gereksinimlerine, sınıf dengesizliğinin yaygınlığına ve farklı sınıflandırma hatası türleri arasındaki takaslara bağlıdır. En yaygın kullanılan metriklerden biri AUC olsa da farklı metrikler farklı performans yönlerini vurgular. Birden fazla metrik kullanmak, sınıflandırıcı performansının daha kapsamlı bir şekilde analiz edilmesine ve daha iyi anlaşılmasına yardımcı olabilir [10].

Sınıflandırma performansını ölçmek için kullanılan metrikler haricinde farklı birçok amaç için kullanılan değerlendirme metriği bulunmaktadır. Örneğin bir model ile keşfedilen biyobelirteçlerin tutarlılığı ve gürbüzlüğü değerlendirilirken yine performans değerlendirme metriklerinden yararlanılır. Sınıflandırma algoritmasının kararlılığını ölçme kavramı, ilk olarak Turney [128] tarafından incelenmiş olup, farklı veri kümeleri üzerinde sınıflandırma kavramları arasındaki uyumun bir ölçüsüne dayalı olarak, kararlılığı niceliksel olarak belirlemek için bir yöntem sunulmuştur. Bir FS algoritmasının kararlılığı, eğitim verilerinin bozulması, parametrelerin farklı ayarları veya algoritmanın farklı rastgele örneklerle başlatılması nedeniyle seçilen öznitelik alt kümesindeki değişimle ilgilidir. Kararlı bir FS algoritması, örneğin hastalıkla ilişkili genleri ayırmak için, alan uzmanına belirli bir güven düzeyi göstermesi gerektiğinden bilgi keşfi için oldukça önemlidir. Farklı eğitim kümelerinin, başlangıç öznitelik kümesindeki yüksek düzeyde benzerlik nedeniyle aynı sınıflandırma sonucuna yol açabilecek farklı öznitelik

alt kümeleri üretmesi mümkündür. Bu durumda kararlı kabul edilebilecek sınıflandırma algoritmasının aksine, öznitelik seçimi algoritması farklı çıktılar üretir. Bu sebeple bir sınıflandırma algoritmasının kararlılık ölçümü kavramı, teknik olarak herhangi bir FS algoritmasının kararlılık ölçümü için kullanılamaz [129]. FS algoritmasının kararlılığı ile ilgili kapsamlı bir analiz üzerine ilk yayınlanmış çalışma Kalousis vd. [130] tarafından 2006 yılında sunulmuştur.

FS algoritmaları genel olarak, sıralanmış veya ağırlıklı bir öznitelik listesinde veya seçilen özniteliklerin optimum bir alt kümesinde öznitelik seçimini gerçekleştirir. FS algoritmalarının sonucunda, öznitelik seçimlerini temsil etmedeki farklılıklara bağlı olarak, kararlılıklarının değerlendirilmesi farklıdır. Buna göre, farklı FS kategorilerinin kararlılığını değerlendirmek için uygun çeşitli kararlılık ölçüleri geliştirilmiştir. Öznitelik alt kümesi seçimi algoritması için çeşitli kararlılık ölçüleri olsa da başta Kuncheva'nın tutarlılık indeksi [131] olmak üzere, benzerliğe dayalı ölçütler oldukça popülerdir.

Kuncheva İndeksi: Kuncheva indeksi (KI) [131], Ludmila Kuncheva tarafından ilk olarak 2007 yılında Sıralı ileri seçim adlı FS algoritması ile gerçekleştirilen deneylerle tanıtılmıştır. FS algoritmaları tarafından belirlenen aday alt kümelerin kalitesinin tahmini, verilerin eğitim/test kümeleri olarak ayrılmasına bağlıdır ve bu alt kümeler farklı öznitelik dizileri döndürebilir. Bu diziler arasındaki önemli tutarsızlıklar, seçimde bir sorun olduğunu göstermektedir bu sebeple KI, kesişimin benzerliğine ve şansa yönelik bir düzeltmeye dayalı bir kararlılık indeksi olarak önerilmektedir. Denklem 2.25' te FS algoritması tarafından belirlenen iki aday alt küme arasındaki Kuncheva benzerlik indeksi hesaplaması verilmiştir. F_i ve F_j , i 'inci ve j 'inci alt örnek üzerinden biyobelirteç saptama yönteminin çıktısını, p orijinal veri kümesindeki toplam öznitelik sayısını belirtir. $n = |F_i| = |F_j|$ olmak üzere, i ve j ' inci alt örnek için Kuncheva benzerlik indeksi aşağıdaki gibi hesaplanır:

$$KI(F_i, F_j) = \frac{p \cdot |F_i \cap F_j| - n^2}{n(p - n)} \quad (2.25)$$

KI, -1 ile 1 aralığında bir değer alabilir. Değer ne kadar büyükse biyobelirteçlerin iki küme arasında o kadar yaygın olduğunu gösterir. Negatif değerler ise seçilen biyobelirteçlerin çoğunlukla rastgele olduğunu gösterir. Algoritmanın genel tutarlılığı

(C_{avg}), tüm ikili alt küme benzerliklerinin ortalamasıdır. m , alt küme sayısı olmak üzere genel tutarlılık performansı aşağıdaki gibi ölçülebilir:

$$C_{avg} = \frac{2 \sum_{i=1}^m \sum_{j=i+1}^m KI(F_i, F_j)}{m(m-1)} \quad (2.26)$$

2.7. Mikrobiyom Verilerinde Sınıflandırma Analizi için Kullanılan Derin Öğrenme Tabanlı Çalışmalar

Mikrobiyom araştırmalarında büyük miktarda veri üretilmesi nedeniyle, DL yöntemleri bu alanda yaygın olarak kullanılmaktadır. Örneğin, MetAML [37], bolluk veya belirteç profilleri üzerinde klasik ML yöntemlerini kullanarak hastalık tahminiyle ilgilenir. MicroPheno [132], 16S rRNA dizilerini alt örnekleme amacıyla bir çarpaz doğrulama tekniği kullanır, daha sonra alt örnekleme dizilerinin k-mer gösterimlerini oluşturarak bu gösterimleri hastalık tahmini için kullanır. MDeep [133], mikrobiyal türler arasındaki filogenetik korelasyonu yakalamak için CNN katmanlarını kullanan ve hem regresyon hem de ikili sınıflandırmalarda üstün performans gösteren bir yöntemdir. ML Repo [134] ise, belirli bir mikrobiyom veri kaynağı olmayan algoritma geliştiricilerin, çeşitli mikrobiyom veri kümeleriyle çalışmasına olanak sağlar. DL tekniklerinin, bağırsak mikrobiyomu ile ilgili çalışmalarda çeşitli hastalıklar üzerindeki tahmin performansını geliştirmek amacıyla sıklıkla kullanıldığı gözlemlenmiştir. Bu tür çalışmaların bir örneği MDL4Microbiome [135]'dur. MDL4Microbiome, sınıflandırma doğruluğunu artırmak için mikrobiyom verilerinin üç farklı kategorisini, yani geleneksel taksonomik profilleri, genom düzeyindeki göreceli bolluğu ve metabolik işlevsel özellikleri birleştirir. Oudah, M. vd. [136] ise yine benzer amaçla, sınıflandırma doğruluğunu artırmak ve sağlık koşullarına ilişkin içgörüler sağlamak için otomatik özellikler kullanan bir algoritma geliştirmiştir.

CNN'ler, mikrobiyom araştırmalarında en yaygın kullanılan DL yöntemleri arasındadır. Geleneksel yaklaşımlar, genellikle farklı taksonomik birimleri bir araya getirmek için taksonomik ağaçların hiyerarşik yapısına güvenir, ancak bu yöntemler bazen, özellikle de nadir taksonlardan gelen verilerin tüm zenginliğini yakalamak konusunda, başarısız olur. Bu sınırlamaların üstesinden gelmek için, mikrobiyom verilerini mikrobiyal taksonlar

arasındaki ilişkileri daha iyi temsil eden görüntülere veya grafiklere dönüştüren Grafik Evrişimli Ağlar (GCN'ler) [137] dahil olmak üzere birçok CNN tabanlı DL tekniği önerilmiştir. Örneğin, TaxoNN [81], OTU'ları kümelere ayırır ve her küme için CNN'leri eğitir. Ardından bir topluluk öğrenme yaklaşımı, tahmin doğruluğunu artırmak için bu CNN'leri birleştirir. CNN'ler, verileri mikrobiyal taksonların mekânsal ilişkilerine göre sınıflandırarak, mikrobiyom araştırmaları dahil olmak üzere birden fazla alanda etkili olduklarını kanıtlamıştır [67, 80, 138]. Örneğin, PopPhy-CNN [67] ve TopoPhy-CNN [138] gibi yöntemler, taksonlar arasındaki ilişkileri korumak için filogenetik ağaçlar oluşturur ve hastalık durumlarını tahmin etmek için CNN'leri kullanır. Bu yöntemler taksonomik yapıyı korurken aynı zamanda tahmin doğruluğunu iyileştirmede başarılı sonuç vermiştir. phyLoSTM [80], CNN'leri ve Uzun Kısa Süreli Bellek Ağlarını (Long Short Term Memory Networks, LSTM) kullanarak insan mikrobiyomundan mekânsal olarak ilişkili verilerin tespitinde tahmin doğruluğunu artırır.

Yüksek boyutlu, düşük örnek içeren verilerdeki problemleri hafifletmek için, verileri daha düşük boyutlu öznitelik uzaylarında modelleme gibi temel ön işleme yaklaşımları mikrobiyom çalışmalarında sıklıkla uygulanmıştır. Araştırmacılar, bu yaklaşımın iyi bilinen bir uygulayıcısı olarak, verileri düşük boyutlu uzayda temsil edebilmek ve öznitelik çıkarımı için popüler bir DL yöntemi olan otokodlayıcıları kullanırlar. Örneğin, Oh M. vd. [139], özellikle kanser immün terapisini iyileştirmek için, mikrobiyom profillerinin genelleştirilebilirliğini ve yorumlanabilirliğini artırmak amacıyla, genelleştirilmiş yorumlanabilir bir otokodlayıcıyı kullanan DeepGeni'yi tanıtmıştır. Öte yandan DeepMicro [85], dört farklı otokodlayıcı kullanarak yüksek boyutlu mikrobiyom verilerini düşük boyutlu bir gösterime dönüştürür ve ardından SVM, RF ve MLP gibi ML sınıflandırma algoritmalarını bu azaltılmış gösterime uygulayarak farklı veri kümeleri arasında model eğitimini ve optimizasyonunu hızlandırır. García-Jiménez vd. [140], mikrobiyal bolluk değerlerini derin bir gizli uzay gösterimine yoğunlaştırmak için otokodlayıcılara dayalı bir YSA kullanarak mikrobiyom kompozisyonlarını tahmin etmeyi amaçlamışlardır. EnsDeepDP [141], derin gösterimlere dayalı bir hastalık puanlama stratejisi geliştirmek için denetimli ve denetimsiz öğrenmeyi birleştirir. Bir diğer yöntem olan MVIB [142] ise, mikrobiyom tabanlı hastalık tahmini için tür göreceli bolluğu ve suş seviyesi belirteç profillerini analiz eder.

GAN, üretici ağ ile ayırıcı ağı çekişmeli eğitim yaklaşımı ile bir araya getiren yenilikçi bir model olarak birçok çalışmaya yön vermiştir. GAN'lar, özellikle son beş yılda mikrobiyom araştırmalarında güçlü bir araç olarak benimsenmiş olup, gelişmiş sentetik veri üretimi ve iyileştirilmiş hastalık tahmini sağlamaktadır. Literatürde bulunan bu tür uygulamaların bir örneği, mikrobiyom tabanlı hastalık tahmininde yüksek performans gösteren GAN-GMHI [143]'dır. GAN-GMHI, mikrobiyom verilerini kullanarak sağlık ve hastalık durumları arasında etkili bir şekilde ayırım yapmaktadır. Bir başka uygulama ise, boylamsal mikrobiyom verisi hesaplaması için GAN'ları kullanan ve zaman serisi analizlerinde daha doğru ve eksiksiz veri kümeleri sağlayan DeepMicroGen'dir [144]. Benzer şekilde, MB-GAN [145], çeşitli alt akış analizlerinde yardımcı olabilecek sentetik veri kümeleri üreterek mikrobiyom verilerini simüle etmek için geliştirilmiştir. MB-GAN, klasik GAN ağına benzeyen bir ağ kullanır. Üretici ağın amacı, rastgele gürültü alıp, bir dizi doğrusal olmayan dönüşüm yapmak ve simüle edilmiş mikrobiyom bolluklarını hesaplamaktır. Ayırıcı ağın amacı ise, verilerin üreticiden mi yoksa gerçek veri kümelerinden mi olduğunu ayırt etmektir. MB-GAN mimarisi, GAN çerçevesine benzerken, mikrobiyom verileri için daha kullanışlı hale getirmek adına 2 özel katman tasarlanmıştır. İlk olarak, filogenetik bir dönüşüm katmanı eklenmiştir. Birçok mikrobiyal takson düşük bolluğa sahip olduğundan, daha iyi model yakınsaması elde etmek için bu düşük değerleri daha büyük değerlere dönüştürmek amaçlanmıştır. İkinci olarak, ayırıcı ağ için kayıp fonksiyonunu hesaplamak amacıyla Earth Mover's (EM) mesafesi kullanılmıştır. Bu sayede, filogenetik bilgileri verilen herhangi iki mikrobiyom örneği arasındaki benzerlik nicel olarak ölçülebilir. Hesaplanan çeşitlilik değeri, ayırıcı ağın kayıp değeri olarak kullanılmıştır. Başka bir çalışmada, Koşullu GAN'lar [146], duruma özgü veriler üreterek, mikrobiyom veri kümelerindeki ML modellerinin performansını artırmak için kullanılmış ve sınıflandırma doğruluğu iyileştirilmiştir. Bu çalışmada, GAN'ın çoklu çeşitlilik ölçütlerine dayalı orijinal verileri temsil eden örnekler oluşturabilmesinin yanı sıra, sentetik örnekler üzerinde ML modellerinin eğitilmesinin, veri artırma yoluyla hastalık tahminini iyileştirebileceğini göstermek için İBH ile ilişkili kohortlar kullanılmıştır. MADGAN [147] ise, hastalık tahmininde mikrop-hastalık ilişkilerini tahmin etmek için GAN'ları kullanır ve karmaşık mikrobiyal etkileşimleri anlamak için yeni bir yaklaşım sunar. Ayrıca, yine GAN tabanlı bir model, hastalık gen tahmini için RNA-seq verilerine de uygulanmıştır [148] ve bu da GAN'ların farklı

biyolojik veri bağlamlarında çok yönlülüğünü göstermektedir.

U-Net modeli biyomedikal görüntü verileriyle sıklıkla çalışılmış olsa da bilinen literatürde henüz bu modelleri kullanarak mikrobiyom verilerinde sınıflandırma analizi yapan ve buradan çeşitli uygulamalar geliştiren bir çalışma bulunmamaktadır.

2.8. Mikrobiyom Verilerinde Biyobelirteç Keşfi için Kullanılan Derin Öğrenme Tabanlı Çalışmalar

Öznitelik seçimi (FS), ölçülecek girdi değişkenlerinin sayısını ve verilerdeki gürültüyü azaltarak aşırı uyum riskini en aza indirir. Ayrıca FS yöntemleri, potansiyel biyobelirteçler olarak hizmet eden metagenomik veri kümelerindeki özniteliklerin bir alt kümesini tanımlamak için önemlidir. Bu sebeple birçok çalışmada, mikrobiyom verilerinde biyobelirteç keşfi için DL tabanlı biyobelirteç tanımlama algoritmaları kullanılmıştır. Örneğin Ditzler vd. [149], Amerikan Bağırsak Projesi ve İBH veri kümeleri üzerinde, Ortak Karşılıklı Bilgi (JMI), RF, Lasso, mRMR ve Karşılıklı Bilgi Maksimizasyonu (MIM) gibi FS yöntemleri kullanılarak bulunan öznitelik alt kümelerini tartışmış ve yöntemlerin değerlendirme sürelerini karşılaştırmıştır. Abbas vd. [34], İBH için metagenomik verilerden güvenilir mikrobiyal biyobelirteçleri belirlemek üzere Rastgele Orman Öznitelik Önemi tarafından seçilen öznitelik örneklerini ve en iyi performans gösteren ağ tabanlı FS çerçevesini birleştiren hibrit yöntem önermiştir. Bu yöntemle seçilen özniteliklerin AUC puanları hesaplanmış ve benzerler yöntemlerden daha iyi performans gösterdiği bildirilmiştir. GDmicro [150], beş farklı hastalık için on kohorttan insan bağırsak mikrobiyomu verileri üzerinde, sınırlı etiketli örnekler kullanarak genelleştirilmiş bir model elde edebilmek için yarı-denetimli DL modelleri kullanmıştır. Sınıflandırma performansı dört farklı popüler yöntemle karşılaştırılmış ve bu yöntemlerden daha iyi performans gösterdiği ve bolluğa dayalı istatistiksel analiz yöntemlerine (Wilcoxon testi) kıyasla kararlı biyobelirteçler bulabildiği bildirilmiştir. Ayrıca, Gou vd. [151], bağırsak mikrobiyomu içinde T2D ile ilgili gürbüz öznitelikleri belirlemek için yorumlanabilir bir ML çerçevesi kullanmıştır. Üç Çin kohortunda tanımlanan 14 özniteliğe dayalı bir mikrobiyom risk puanı (MRS) hesaplayarak, MRS ile glikoz seviyelerindeki artışlar ve bağırsak mikrobiyotasından türetilen çeşitli metabolitler arasında bir bağlantı kurmuştur. Ek olarak, Larsen vd. [152], mikrobiyom etkileşiminden

disbiyozu tahmin etmek için ML tekniklerini kullanmış ve insan mikrobiyomu ile sağlık arasındaki bağlantılara ilişkin potansiyel içgörüler sağlamayı amaçlamıştır. Sivakumar vd. [153] çalışmalarında, derin ileri beslemeli sinir ağına (DFNN) yerleştirilmiş MAGMA kullanarak seyrek bir mikrobiyal etkileşim ağının (MIN) oluşturulması yoluyla, grafik gömme FS (Graph embedding feature selection) yaklaşımını araştırmıştır. Yöntem, deneysel olarak farklı sınıflandırıcılar ve İBH ve KK veri kümeleri için DFNN'ye yerleştirilmiş farklı MIN modelleri ile çapraz doğrulamaya tabi tutulmuş, ayrıca seçilen biyobelirteçler İBH ve KK veri kümeleri için biyolojik çalışmalarla kıyaslanmıştır. Le vd. [4], mikrobiyal bolluklardan metabolit bolluğunu tahmin etmek için bir kodlayıcı-kod çözücü sinir ağı çerçevesi sunmuştur. Sunulan modelde ağın gizli katmanı, klinik olarak anlamlı olan temel mikrop-metabolit ilişkilerini yakalar. Alshawaqfeh vd. [1], karmaşık mikrobiyal etkileşimleri göz önünde bulundurarak, metagenomik veri matrisini iki ayrı matrisin çarpımı olarak modellemiş ve seyrek matristen gürbüz biyobelirteçleri tespit edebilen RPCA adlı bir yöntem geliştirmiştir. Yöntemin iki gerçekçi veri kümesi üzerinde güncel algoritmalarla kapsamlı karşılaştırma analizleri gerçekleştirilmiş ve RPCA'nın sınıflandırma doğruluğu ve tekrarlanabilirliği açısından diğer algoritmalarından daha iyi performans gösterdiği raporlanmıştır.

U-Net ve GAN tabanlı modeller, görüntü verileriyle birçok kez çalışılmış olsa da bilinen literatürde, mikrobiyom verilerinde biyobelirteç keşfi için bu modelleri kullanan bir çalışma şuan için bulunmamaktadır.

3. BÖLÜM

MİKROBİYOM VERİLERİNDE SINIFLANDIRMA PERFORMANSININ GELİŞTİRİLMESİ

3.1. DeepGum: Mikrobiyom Analizi için Derin Öznitelik Aktarımı

3.1.1. Genel Bakış

Mikrobiyom verilerinin yüksek boyutlu, az örnekli yapısının DL modelleri için önemli bir zorluk sunduğu açıktır. Bu sorunların üstesinden gelmek ve az sayıda örneğe sahip mikrobiyom veri kümelerinde yüksek doğrulukta sınıflandırma sonuçları elde edebilmek amacıyla, bağırsak mikrobiyom analizi için derin öznitelik aktarımı uygulamasını içeren DeepGum isimli yeni bir yaklaşım geliştirilmiştir. DeepGum, yüksek boyutlu, düşük örneklemlili bağırsak mikrobiyomu veri kümelerinde, maksimum verimlilikle öznitelikler çıkarmak için denetimsiz DL modellerini kullanır. Bu çalışmada, literatürdeki tek bir veri kümesinde boyutsallığı azaltmanın geleneksel kullanımının aksine, darboğaz (bottleneck) modellerine, çok kohortlu bir metagenom koleksiyonunda temsili öğrenenler olarak yaklaşılmıştır. Birden fazla veri kümesinde modelleme yapılırken, transfer öğrenme ile tek bir görev için ince ayar yapılmış ve sınırlı sayıda örneğe sahip veri kümelerinde bile dikkate değer sınıflandırma performansı elde edilmiştir.

3.1.2. Çalışmada Kullanılan Mikrobiyom Veri Kümeleri ve Ön İşleme

Başlangıçta, kapsamlı bir bağırsak mikrobiyomu veri kümesi oluşturabilmek amacıyla, ENA moleküler veri tabanından [154] GI hastalıklarla ilgili dokuz farklı veri kümesi temin edilmiştir. Çeşitli araştırma makalelerinden toplanan bu dokuz veri kümesinden, beş veri kümesi (İBH-I, İBH-II, İBH-III, İBH-IV ve İBH-V) İBH kohortlarından

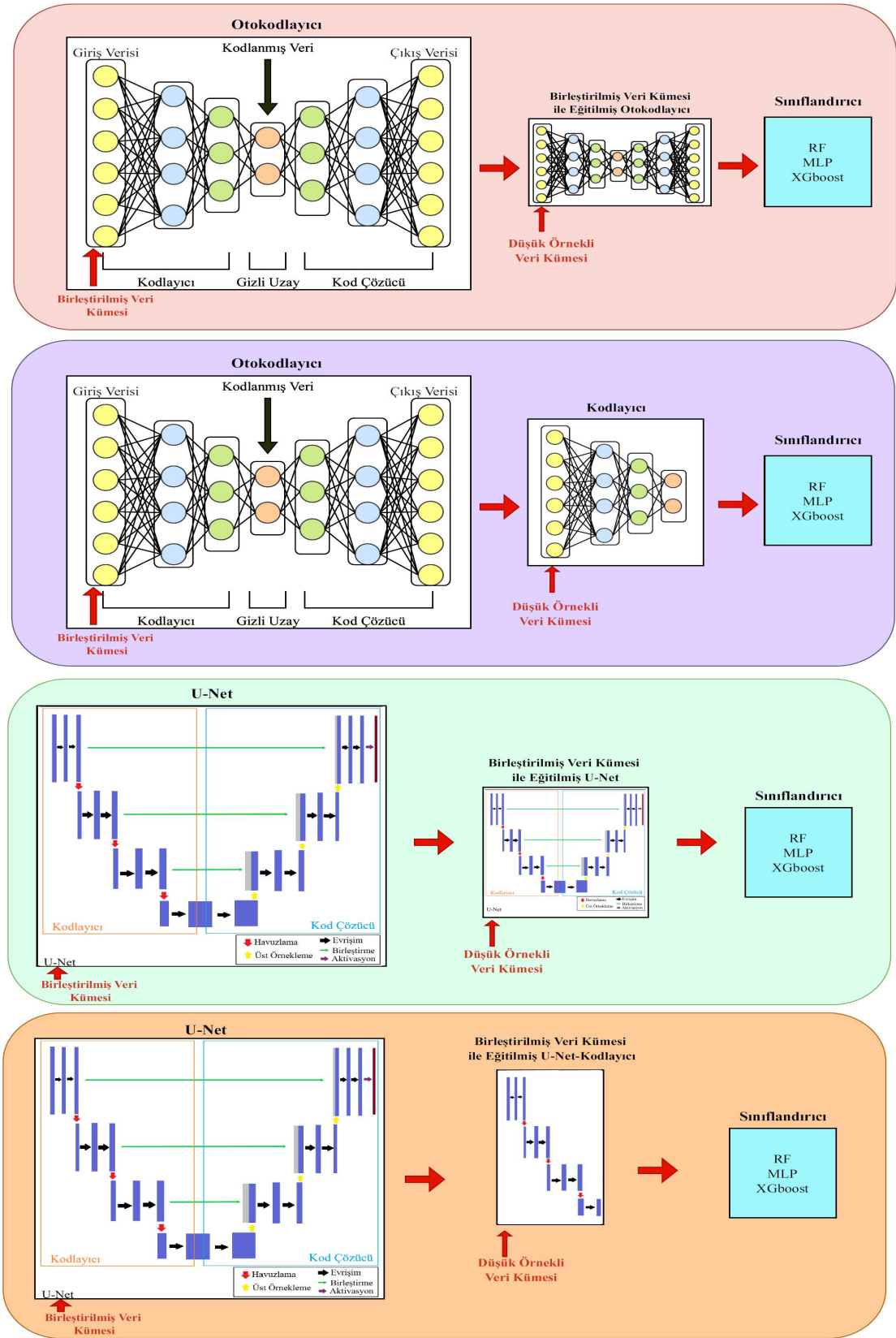
Tablo 3.1. Bağırsak mikrobiyomu veri kümelerine ait detaylı bilgiler.

Veri Kümesi	Toplam Örnek Sayısı	Hasta Örnekleri	Sağlıklı Kontrol Sayısı	Etiketsiz Örnek Sayısı
İBH-I	213	ÜK=72, CH=85	56	*
İBH-II	335	ÜK=88, CH=170	56	21
İBH-III	539	ÜK=242, CH=69	228	*
İBH-IV	426	*	*	426
İBH-V	338	*	*	338
KK-I	155	46	66	43
KK-II	130	70	60	*
KK-III	118	53	53	*
KK-IV	123	*	*	123

oluşurken [39, 155–158], kalan dört veri kümesi (KK-I, KK-II, KK-III ve KK-IV) ise KK kohortlarını içermektedir [159–162]. Bu veri kümelerini tanımlamak için çalışmada kullanılan terminoloji, Tablo 3.1’de, veri kümesi kısmında verilmiştir. Ayrıca veri kümelerine ait veri dağılımı, toplam örnek sayısı ve etiket bilgisi gibi detaylı bilgiler de bu tabloya eklenmiştir. Elde edilen veri kümeleri MetaPhlAn 3.0 [163] ile işlenerek taksonomik öznitelikleri çıkarılmıştır. Daha sonra veri kümeleri ortak taksonomik özniteliklerine göre birleştirilerek, toplamda 2377 örnek ve 1104 taksonomik öznitelik içeren heterojen birleştirilmiş bir veri kümesi üretilmiştir. Taksonomik birimler, tablo gösteriminde benzer taksonları mekânsal olarak yakın şekilde gruplamak için leksikografik olarak sıralanmıştır. Ayrıca bu sıralama ile, CNN modellerinin kısa vadeli korelasyonları yakalamasına yardımcı olmak amaçlanmıştır.

3.1.3. Denetimsiz Yöntemlerle Derin Öznitelik Çıkarımı

Bu çalışmada, az sayıda örneğe sahip mikrobiyom veri kümelerinde yüksek doğrulukta sınıflandırma sonuçları elde edebilmek amacıyla temel olarak iki farklı DL yaklaşımından (Otokodlayıcı ve U-Net) yararlanılır. Otokodlayıcı modelin iki ana rolü vardır (Şekil

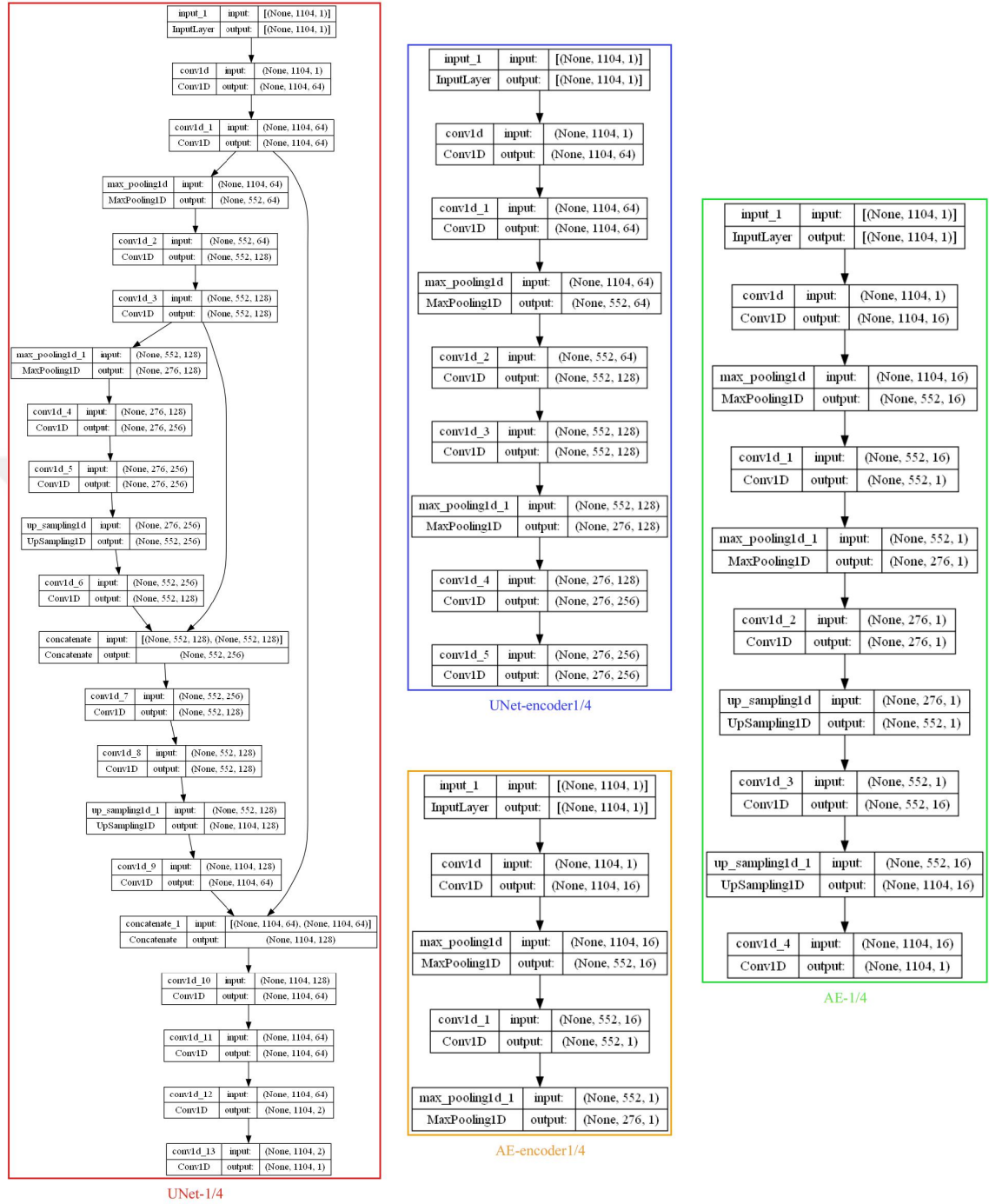


Şekil 3.1. DeepGum yönteminin genel mimarisi. Bu yaklaşım, birleştirilmiş mikrobiyom veri kümesiyle bir Otokodlayıcı veya U-Net'i eğitmeyi içerir. Eğitim yoluyla

ortaya çıkan kodlayıcı model veya yeniden yapılandırılmış model, az örneğe sahip veri kümeleri için tekrar kullanılır. Daha sonra, örneklerin hastalara mı yoksa sağlıklı kontrollere mi ait olduğunu belirlemek için bir sınıflandırıcı model (MLP, XGB veya RF) kullanılır.

3.1). İlk olarak, birleştirilmiş mikrobiyom veri kümesinin gizli bir temsilini oluşturmak ve bir ‘mikrobiyom alanı’ öğrenmek için kullanılır. İkinci olarak, birleştirilmiş veri kümesiyle eğitilen ve gizli alandan minimum kayıpla yeniden yapılandırılan otokodlayıcının çıktısı kullanılır. Bu kodlayıcı modeli ve yeniden yapılandırılmış model daha sonra etiketli, küçük örneklemli veri kümelerinde (İBH-I, İBH-II, İBH-III, KK-I, KK-II ve KK-III) hastalık fenotipi tahmini gayesiyle sınıflandırıcıları eğitmek için kullanılır. Her biri farklı sayıda örneğe, ancak eşit sayıda öznitelige sahip olan bu altı ayrı veri kümesi, DeepGum’ın transfer öğrenme yeteneklerini sergilemek için seçilmiştir. Kalan üç veri kümesi (İBH-IV, KK-IV, İBH-V), niteliklerinin sınıflandırma yöntemlerinde karşılaştırmalı analiz için uygun görülmemesi nedeniyle değerlendirmeden hariç tutulmuştur. Özetle, bu veri kümelerinde yeterli sayıda sağlıklı kontrol bulunmaması, onları dengesiz ve sınıflandırma performansı kıyaslaması için uyumsuz hale getirmiştir. Hem kodlayıcı hem de kod çözücüde geleneksel tam bağlı katmanlar (FCNN) yerine, Leaky Rectified Linear Unit (LeakyReLU) aktivasyon fonksiyonuna sahip, aynı sayıda, tek boyutlu Evrişimli Sinir Ağı (1B-CNN) kullanılmıştır. Kodlayıcı, giriş verilerinde evrişimler kullanır ve daha sonra maksimum havuzlama yoluyla boyutsallığı azaltarak verimli bir öznitelik gösterimi oluşturur. Kod çözücü ise bu gösterimi yukarı örnekleyerek bu işlemi tersine çevirir. Son olarak, çıktıyı üretmek için bir sigmoid aktivasyonu eklenir. Yeniden yapılandırma kaybını hesaplamak için Ortalama Karesel Hata (MSE) ve optimizasyon işlemi için Adam algoritması kullanılmıştır. Adam optimizasyon algoritması için öğrenme oranı varsayılan değer olan 0,001 olarak ayarlanmıştır.

Çalışmadaki otokodlayıcı modelleri, gizli birim sayısının kodlayıcı bölümünde bir önceki katmanın yarısı, kod çözücü bölümünde ise bir önceki katmanın iki katı olduğu belirli bir tasarıma sahiptir. Bu yaklaşım, modelin yapısındaki simetriyi korurken hem gizli birim sayısını hem de gizli katmanları ayarlayarak model karmaşıklığının kontrol edilmesini sağlar. Örneğin, gizli katman 276 gizli birimden oluşuyorsa, ortaya çıkan otokodlayıcı



Şekil 3.2. DeepGum katmanlarının detaylı bilgileri.

katmanlarındaki birim sayıları Şekil 3.2’de gösterildiği gibi, 1104-552-276-552-1104 olur. Başlangıçta, otokodlayıcı için girdi olarak 1104 özniteliğe sahip birleştirilmiş genel mikrobiyom veri seti kullanılmıştır. Daha sonra, otokodlayıcıdaki gizli gösterim için üç farklı boyutta deneyler yapılmış (276 (orijinal boyutun 1/4’ü), 138 (orijinal boyutun 1/8’i) ve 69 (orijinal boyutun 1/16’sı), sonuç olarak, üç yeniden yapılandırılmış model (AE-1/4, AE-1/8 ve AE-1/16) ve üç kodlayıcı (AE-encoder1/4, AE-encoder1/8 ve

AE-encoder1/16) model elde edilmiştir.

Otokodlayıcı gibi, U-Net mimarisinin ilk bölümü bir kodlayıcı olarak işlev görür ve ikinci bölüm bir kod çözücü olarak hizmet eder. Bu çalışmada, otokodlayıcıya benzer şekilde, U-Net modelinin iki ana amacı vardır: birleştirilmiş mikrobiyom veri kümesinin yeniden yapılandırmak ve onun gizli temsilini oluşturmak. Kodlayıcı model ve yeniden yapılandırılmış model daha sonra sınırlı sayıda örneğe sahip, altı etiketli veri kümesinde hastalık tahmini amacıyla sınıflandırıcıları eğitmek için kullanılır. U-Net'in hem birinci hem de ikinci kısmı, Şekil 3.2'de gösterildiği gibi eşit sayıda 1B-CNN kullanır. U-Net modellerinde, evrişimli katmanlarda aktivasyon fonksiyonu olarak LeakyReLU, çıkış katmanında ise sigmoid aktivasyon fonksiyonu kullanılır. Ayrıca, optimizasyon fonksiyonu olarak Adamax ve kayıp fonksiyonu olarak ise MSE kullanılmıştır. Başlangıçta, 1104 öznelikten oluşan birleştirilmiş genel mikrobiyom verileri U-Net için girdi olarak kullanılır, daha sonra U-Net'teki gizli temsil için üç farklı boyutta (276(1/4), 138(1/8), 69(1/16)) ayrıntılı analiz gerçekleştirilir. Böylece, üç yeniden yapılandırılmış model (UNet-1/4, UNet-1/8 ve UNet-1/16) ve üç kodlayıcı (UNet-encoder1/4, UNet-encoder1/8, UNet-encoder1/16) model elde edilir.

Modellerin sınıflandırma performansı ve yeniden yapılandırma kaybını adil bir şekilde değerlendirme amacıyla, birleştirilmiş veri kümesi eğitim ve test olarak ayrılır ve her model için 8:2 oranını korunur. Modeller 100 döngü (epoch) ve 16 parti (batch) boyutu ile eğitilir. Otokodlayıcı veya U-Net modeli eğitildikten sonra, kodlayıcı ve kod çözücü ağırlıkları, alt akış sınıflandırma aşamalarında herhangi bir ince ayar yapılmadan dondurulur. Modellerin sınıflandırma etkinliğini değerlendirmek için Random Forest (RF), eXtreme Gradient Boosting (XGB) ve Multi-Layer Perceptron (MLP) olmak üzere üç ayrı sınıflandırıcı kullanılmıştır. Bu modeller için çeşitli hiperparametre kombinasyon analizleri yapıldıktan sonra: RF sınıflandırıcının ormandaki ağaç sayısı (tahmin ediciler) 900 ve dahili bir düğümü bölmek için gereken minimum örnek sayısı 2; XGB sınıflandırıcı için gerçekleştirilecek yükseltme aşaması sayısı 900; son olarak, MLP sınıflandırıcı için gizli katman boyutu 100, aktivasyon fonksiyonu ReLu ve maksimum yinleme sayısı 1000 olarak ayarlanmıştır. Sağlam bir performans değerlendirmesi gerçekleştirmek adına, tabakalı 5 katlı çapraz doğrulama (stratified 5fold-CV) gerçekleştirilmiştir. Modellerin performansını karşılaştırmak için, beş ayrı

deneyden elde edilen test sonuçlarının ortalaması kullanılır. ROC AUC değerlendirme metriği kullanılarak sınıflandırma tahmin performansı değerlendirilir.

3.1.4. Örneklem Büyüklüğüne Dayalı Sınıflandırma Analizi

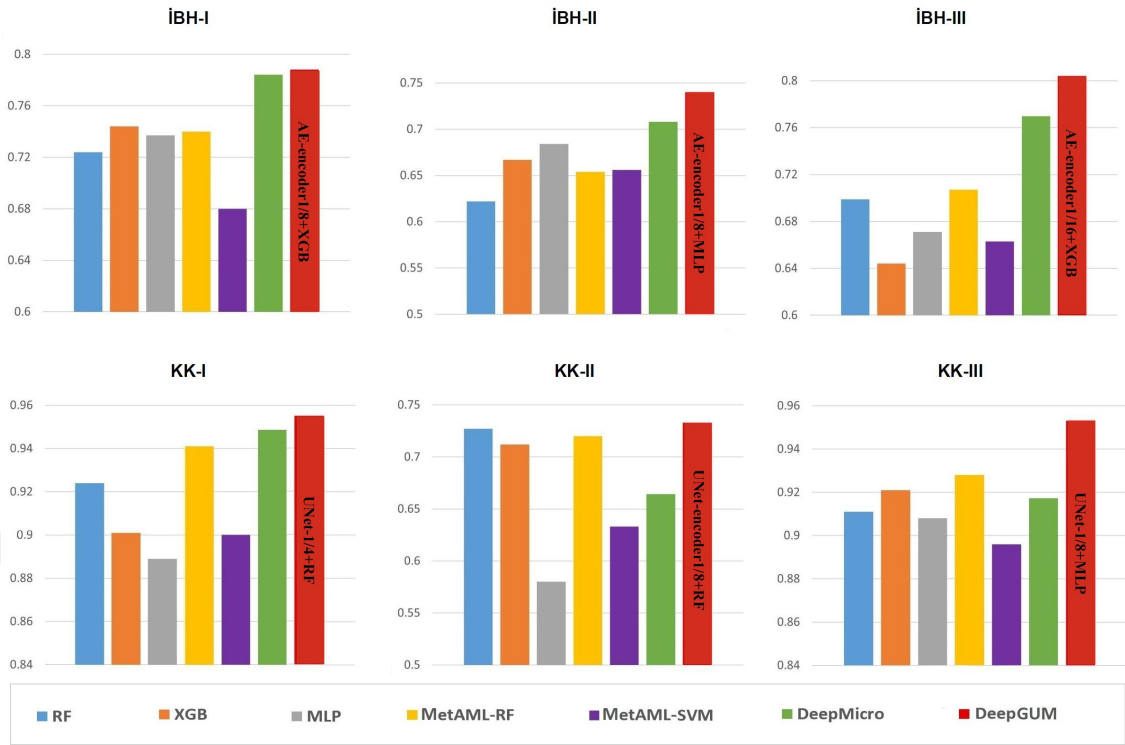
Bu bölümde, yalnızca birkaç örnek sayısı ile insan bağırsak mikrobiyomu örnek etiketini tahmin etme görevi ele alınmıştır. Başlıca amaç, özellikle sınırlı mikrobiyom örnekleriyle (sıklıkla, örnek sayısı <100) uğraşırken yeniden yapılandırma ve gizli alan metodolojilerimizin ne kadar etkili olduğunu değerlendirmektir. Çünkü bu gibi durumlarda, verilerin sınırlı kullanılabilirliği nedeniyle sıfırdan bir tahmin modeli oluşturmak genellikle zor bir görev olmaktadır. Bu sebeple, transfer öğrenimiyle ilgili deneylerimizde, geliştirilen modelin derin gizli alanında depolanan bilgiye erişilmiştir. Bu alan, çok sayıda örneği olan birleştirilmiş bir veri kümesi kullanılarak oluşturulmuştur. Amaç, aktarılan yapılara atfedilebilecek sınırlı sayıda örneği olan veri kümeleriyle uğraşırken bile daha iyi sınıflandırma doğruluğu elde etmektir.

3.1.5. DeepGum Yöntemiyle Elde Edilen Sınıflandırma Analizi Sonuçları

3.1.5.1. Bağırsak Hastalıkları Veri Kümelerinde DeepGum ile Elde Edilen Sınıflandırma Sonuçları

Geliştirilen DeepGum yönteminin sınıflandırma performansı, bağırsak mikrobiyomundan hastalık tahmini için mevcut literatürdeki yöntemlerle, yani RF, XGB, MLP, MetAML [66] ve DeepMicro [85] ile karşılaştırılmıştır. Şekil 3.3'te altı GI hastalık veri kümesi (İBH-I, İBH-II, İBH-III, KK-I, KK-II ve KK-III) üzerindeki performans değerlendirme analizi için, beş deneyin ortalaması ile elde edilen AUC puanları raporlanmıştır.

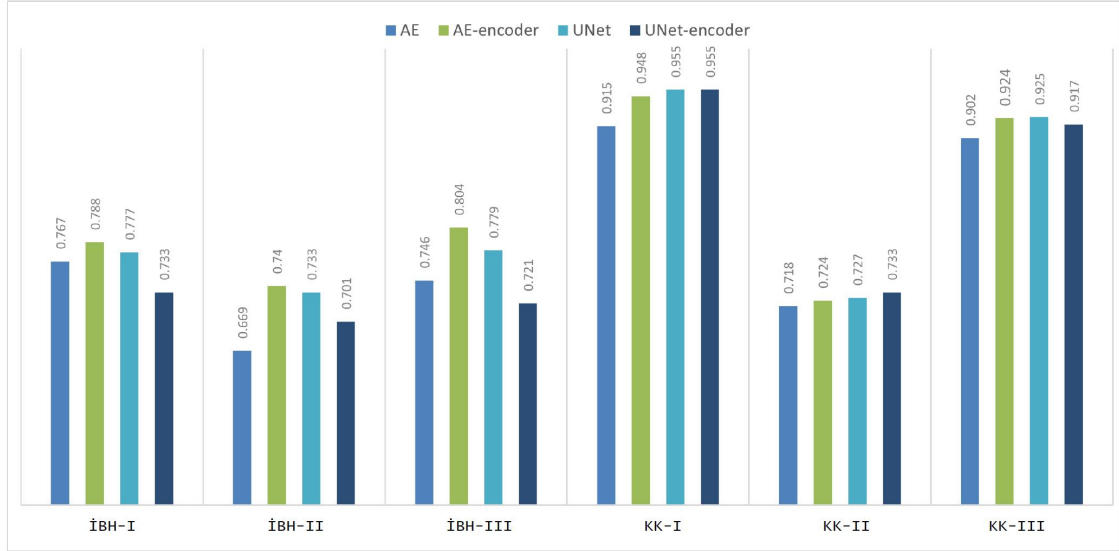
Sonuçlar, hastalık ve kontrol örnek grupları arasında ayırım yapmak için ikili (binary) sınıflandırma sonuçlarını göstermektedir. Bu analiz için, birden fazla hastalık sınıfına sahip (ÜK, CH) veri kümeleri (İBH-I, İBH-II, İBH-III), tüm veri kümelerinde ikili sınıflandırma yapmak amacıyla tek bir hastalık sınıfında birleştirilmiştir. Öncelikle, her DeepGum modeli için, 1104 öznitelik ve 2377 örnek içeren birleştirilmiş veri kümesi, %20 test verisi ve %80 eğitim verisi olacak şekilde bölünür. Test verileri bu süreç boyunca



Şekil 3.3. DeepGum’ın GI hastalıklar için tahmin performansı. IBH-I, IBH-II, IBH-III, KK-I, KK-II ve KK-III veri kümeleri için RF, XGB, MLP, metAML ve DeepMicro modelleri ile Otokodlayıcı ve U-Net tabanlı DeepGum modellerinin karşılaştırmalı sınıflandırma sonuçları.

ayrı tutularak, eğitim verilerine özelleştirilmiş otokodlayıcı veya U-Net modelleri uygulanır. Bu süreçten elde edilen kodlayıcı veya yeniden yapılandırılmış modellerin sınıflandırma performansını, sınırlı örnek içeren veri kümelerinde değerlendirmek için üç ayrı sınıflandırıcı (RF/MLP/XGB) kullanılır. Sınıflandırma performansının sağlam bir şekilde değerlendirilmesini sağlamak için, katmanlı 5 katlı çapraz doğrulama yöntemi kullanılır. Çapraz doğrulamadan elde edilen AUC puanlarının ortalaması alınarak modellerin performansını değerlendirmek ve karşılaştırmak için kullanılır. MetAML ve DeepMicro yöntemlerine ait deneyler için, ilgili orijinal çalışmalarında belgelenen en verimli sınıflandırma sonuçlarını gösteren modeller uygulanmış ve bu çalışmalarda belirtilen özdeş parametreler benimsenmiştir. Detaylandırmak gerekirse, MetAML çalışması için MetAML-RF ve MetAML-SVM modelleri, DeepMicro çalışması için ise CAE-SVM modeli kullanılmıştır. Her veri seti için en iyi performansı gösteren DeepGum modelleri Şekil 3.3’te kırmızıyla vurgulanmıştır. Örneğin, "AE-encoder1/8" gizli gösterimde $1104/8 = 138$ düğüme sahip otokodlayıcı mimarisini belirtir. En iyi sınıflandırma algoritması öğrenilen gösterime göre farklılık göstermiştir. Özetle,

DeepGum dahil olmak üzere çeşitli sınıflandırıcıların ayırım yetenekleri, farklı veri setleri arasında değişmektedir.

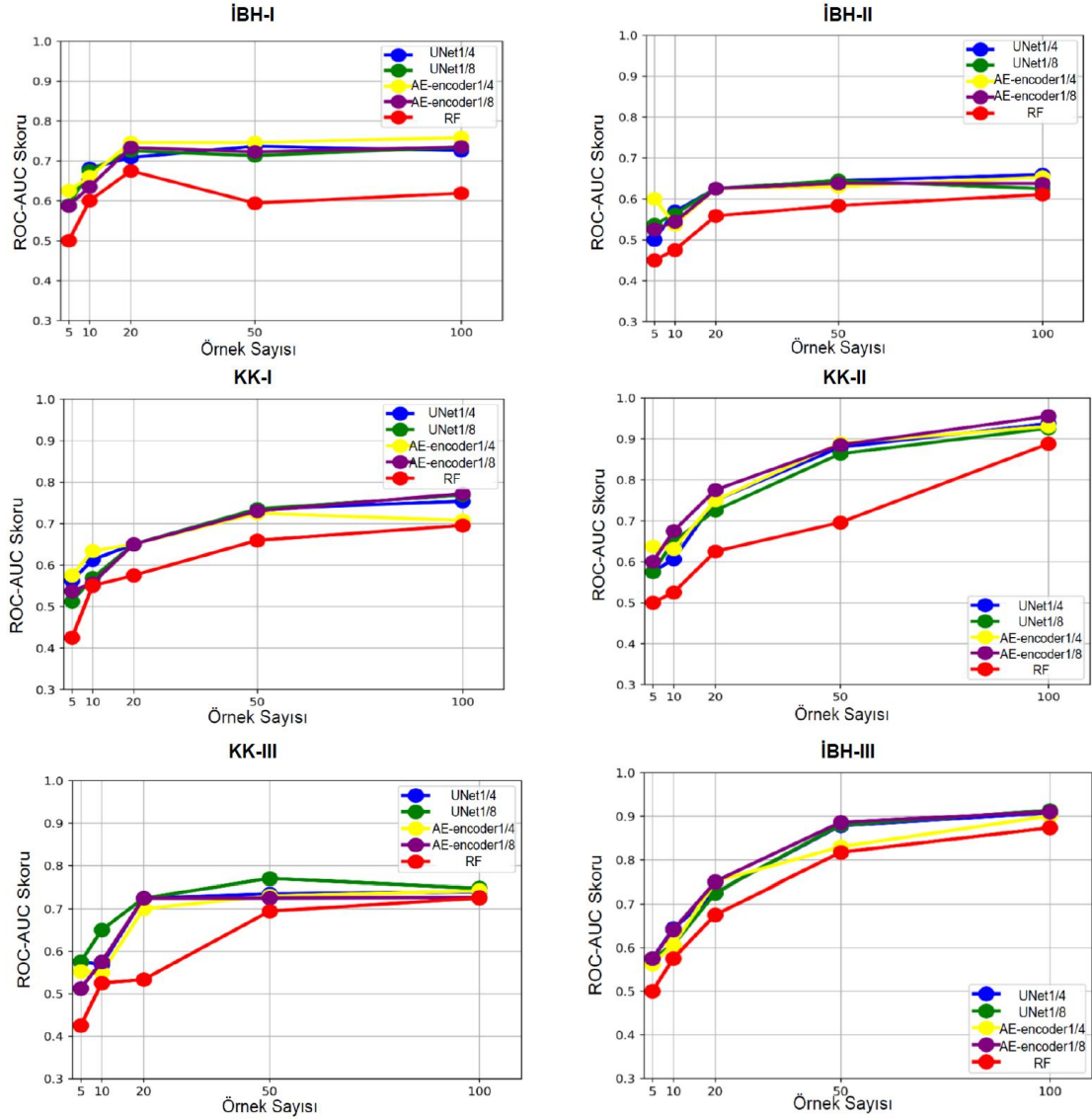


Şekil 3.4. U-Net ve Otokodlayıcı modellerinin yeniden yapılandırılmış (AE, UNet) ve kodlayıcı (AE-encoder, UNet-encoder) modelleri için GI hastalıkları tahmin performansının (AUC) karşılaştırılması.

Her veri kümesi için Otokodlayıcı ve U-Net modelleri değerlendirilmiş ve sonuçlar Şekil 3.4'te sunulmuştur. Farklı renklerle gösterilen her model, en iyi sınıflandırma sonucunu elde eden sınıflandırıcı ile birlikte sunulmuştur. Sonuçlara göre, çoğu durumda kodlayıcı modellerin veri kümeleri için en iyi sonuçları verdiği gözlemlenebilir. Tüm sınıflandırıcılar için en iyi sonucu veren tek bir mimari belirlenememiş, ancak AE-encoder1/4 ve AE-encoder1/8 mimarileri veri kümelerinin çoğunda en iyi performansı göstermiştir. Otokodlayıcı ve U-Net modelleri, önemli bilgileri kompakt bir biçimde yakalamaları dolayısıyla kullanılmıştır. Otokodlayıcılar öğrenme sırasında hangi giriş özelliklerinin kodlanacağına öncelik verir. Doğru model seçiminin genellikle daha iyi bir temsile ve böylece gelişmiş tahmine yol açtığı söylenebilir. Ancak, uygun Otokodlayıcı veya U-Net'in seçimi büyük ölçüde problemin karmaşıklığına ve veri kümesindeki mikrobiyom bileşimindeki farka bağlıdır. Ayrıca, veri kümelerindeki sağlıklı kontrol ve hasta örneklerinin sayısı ve bunların oranı da modellerin verileri ne kadar iyi sınıflandırabileceğini etkiler.

3.1.5.2. Rastgele Seçilmiş Seyrek Örnekli Kohortlar için Sınıflandırma Analizi

Sonuçları



Şekil 3.5. Her veri setinden rastgele seçilen 5, 10, 20, 50 ve 100 örnek için, DeepGum+RF ile belirlenen ve RF sınıflandırıcı model ile belirlenen ortalama AUC puanlarının karşılaştırmalı sonuçları.

Son olarak, DeepGum yaklaşımının sınıflandırma performansı, veri kümeleri içerisinde seçilmiş rastgele az sayıda örnek için değerlendirilmiştir. Bu deneyde, “Önceden eğitilmiş modeller, özellikle birkaç örnek içeren çalışma gruplarıyla uğraşırken sınıflandırma performansını iyileştirebilir mi?” sorusunun yanıtı aranmıştır. Temsili öğrenmenin gücü, bu gibi durumlarda kritiktir. Hedefimiz, denetimsiz ön eğitim yoluyla, sınırlı örneklerle etiketler bulmak için büyük miktardaki etiketlenmemiş verinin gücünden yararlanmaktır. Bu bağlamda, önce modeller çok sayıda örnek içeren birleştirilmiş bir veri kümesiyle

eğitilmiş, daha sonra her veri kümesinden (İBH-I, İBH-II, İBH-III, KK-I, KK-II ve KK-III) rastgele az sayıda örnek (5, 10, 20, 50 ve 100) seçilmiştir. Sonuçların güvenilirliğini sağlamak için bu işlem 5 ve 10 örnek için 40 kez ve 20, 50 ve 100 örnek için 10 kez tekrarlanmıştır. Seçilen örnekler DeepGum modelleri kullanılarak eğitilip, RF sınıflandırıcısı ile sınıflandırıldıktan sonra ortalama AUC değerleri hesaplanmıştır. DeepGum modellerinde, 5 ve 10 örnek için test boyutu %30 seçilirken, 20, 50 ve 100 örnek için %20 oranında test boyutu kullanılmıştır. Veri kümeleri, eğitim ve test kümeleri arasında, aynı sınıf dağılımı oranını koruyacak şekilde bölünmüştür. Aynı koşullar karşılaştırma için seçilen RF sınıflandırıcısı için de uygulanmıştır. Bu deneylerden elde edilen ortalama AUC değerlerinin karşılaştırmalı sonuçları Şekil 3.5'te sunulmuştur. Objektif bir değerlendirme yapmak amacıyla, her veri kümesi için aynı dört modelin (AE-encoder1/4, AE-encoder1/8, UNet-1/4 ve UNet-1/8) sınıflandırma sonucu, görsel olarak kırmızı renkle gösterilen RF sınıflandırıcı çıktılarıyla karşılaştırılmıştır.

3.2. Deeptaxim: Taksonomik Veri Kümeleri için Görüntü Tabanlı Derin Öğrenme Modelleri ile Kapsamlı Sınıflandırma Analizi

3.2.1. Genel Bakış

Mevcut literatürdeki boşluklar göz önünde bulundurularak, DL modellerinin yardımıyla yüksek boyutlu, seyrek mikrobiyom verilerinde doğru ve güvenilir sınıflandırma için yeni bir yaklaşım geliştirilmiştir. Bağırsak mikrobiyomu değişiklikleriyle güçlü ilişkileri olduğu bilinen hastalıklara odaklanılmış ve mikrobiyom verileri [164]'ye benzer şekilde taksonomik ağaç (kladogram) hiyerarşik yapısı kullanılarak görüntü verisi olarak işlenmiştir. Otokodlayıcı, U-Net ve GAN modelleri dahil olmak üzere, DL modelleri kullanılarak geliştirilen modelin (bundan sonra Deeptaxim olarak adlandırılacaktır) performansı iki ayrı veri kümesi grubunda test edilmiştir. Bu yaklaşımın temel amacı dört maddede açıklanabilir: (1) mikrobiyom verilerini, kladogramları kullanarak görüntü verilerine dönüştürmek ve bu görüntüleri CNN tabanlı yöntemlerle eğiterek daha fazla içgörü elde etmek, (2) mikrobiyom tabanlı hastalık sınıflandırmasının doğru bir şekilde yapılması için DL modelleri geliştirmek ve optimize etmek, (3) birleştirilmiş veri kümesi ile eğitilen Deeptaxim'in düşük örnekli veri kümelerinde sınıflandırma performansını

iyileştirmek için transfer öğrenme kapasitesini ölçmek ve (4) Deeptaxim modelinin daha geniş bir hastalık kümesine aktarılabilirliğini değerlendirerek genel uygulama için sağlamlığını ve potansiyelini test etmek.

3.2.2. Kullanılan Veri Kümelerinin Detayları ve Ön İşleme

Deeptaxim için iki farklı veri kümesi grubu ile çalışılmıştır. Başlangıçta bağırsak mikrobiyomu disbiyozu ile bilinen ilişkileri olan İBH ve KK hastalıklarını içeren ilk veri seti grubu (Gastro veri kümesi) ile deneyler gerçekleştirilmiştir. ENA moleküler veri tabanı [154] aracılığıyla erişilen ve çeşitli araştırma makalelerinde kamuya açık olarak yayınlanan dokuz veri kümesi için hasta örneklerinin sayısı, sağlıklı kontrol örneklerinin sayısı, toplam örnek sayısı, veri setlerine verilen isimler Tablo 3.1’de verilmiştir. Veri kümelerinden taksonomik öznitelikleri çıkarmak amacıyla MetaPhlAn 3.0 [163] kullanılmış, ardından veri setleri ortak taksonomik özniteliklerine göre birleştirilerek toplam 2377 örnek ve 1104 taksonomik öznitelik içeren heterojen bir şekilde birleştirilmiş bir veri kümesi üretilmiştir.

Ardından, geliştirilen yöntemlerin sağlamlığını ve genelleştirilebilirliğini doğrulamak

Tablo 3.2. Çoklu-Hastalık veri kümesine ait detaylı bilgiler.

Fenotip	Örnek Sayısı
Sağlıklı	5547
ACVD	214
AS	95
KK	789
CH	284
GH	100
KS	152
MS	24
NAYKH	86
RA	151
T2D	377
ÜK	250

için, analizler daha geniş bir hastalık yelpazesini kapsayan ve daha fazla örnek içeren ikinci bir veri kümesi grubuyla (Çoklu-Hastalık veri kümesi) genişletilmiştir. Bu veri kümesinde ele alınan hastalıklar yalnızca GI bozuklukları değil, aynı zamanda bağırsak mikrobiyomu değişiklikleriyle bağlantılı olan metabolik, kardiyovasküler ve nörodejeneratif hastalıkları da içermektedir. "Bağırsak Mikrobiyomu Sağlık İndeksi 2 (GMWI2)" çalışmasıyla [165] uyumlu ikinci veri kümesi grubu, Aterosklerotik kardiyovasküler hastalık (ACVD), Ankilozan spondilit (AS), Kolorektal kanser (KK), Crohn hastalığı (CH), Graves hastalığı (GH), Karaciğer Sirozu (KS), Multipl skleroz (MS), Non-Alkolik Yağlı Karaciğer Hastalığı (NAYKH), Romatoid Artrit (RA), Tip-2 diyabet (T2D) ve Ülseratif kolit (ÜK) olmak üzere 11 farklı hastalıktan örnekler içerir. Çoklu-hastalık veri kümesi olarak isimlendirdiğimiz bu ikinci grup, toplamda 8069 örnek ve 1211 özniteliğe sahiptir. Detaylı bilgiler Tablo 3.2'de verilmiştir.

3.2.2.1. Mikrobiyom Verilerinden Görüntü Verilerine Geçiş Süreci

Yetersiz zenginliğe sahip mikrobiyomlar (%5'ten yüksek oranda sınıflandırılmamış okumaya sahip) ileri analizlerden hariç tutulmuştur. Ayrıca, örneklerin %1'inden düşük bir kısmında gözlemlenen taksonları ortadan kaldıran bir yaygınlık (prevalence) filtresi kullanılmıştır. Çalışma grubundaki taksonomik verilerinin ön işlenmesinin ardından, taksonomik vektörler MIPMLP [166] boru hattı kullanılarak dönüştürülmüştür. MIPMLP içindeki bu ön işleme dört aşamayı kapsar: taksonomiye dayalı benzer öznitelikleri birleştirme, dağılımı ölçekleme, z puanlarına göre standartlaştırma ve boyut azaltma. Öznitelikler, Sub-PCA kullanılarak tür taksonomisi düzeyinde birleştirilmiş, log normalizasyonu ve z puanlaması atlanmıştır. Her takson için, her girişin belirli bir taksonomik düzeyde bir mikrop temsil ettiği 8 boyutlu bir vektör oluşturulmuştur. [164]'de açıklandığı gibi, hiyerarşik kümeleme yoluyla taksonların yeniden sıralanmasından sonra, her biri bir metagenom örneğini temsil eden taksonomik görüntüler oluşturulmuştur.

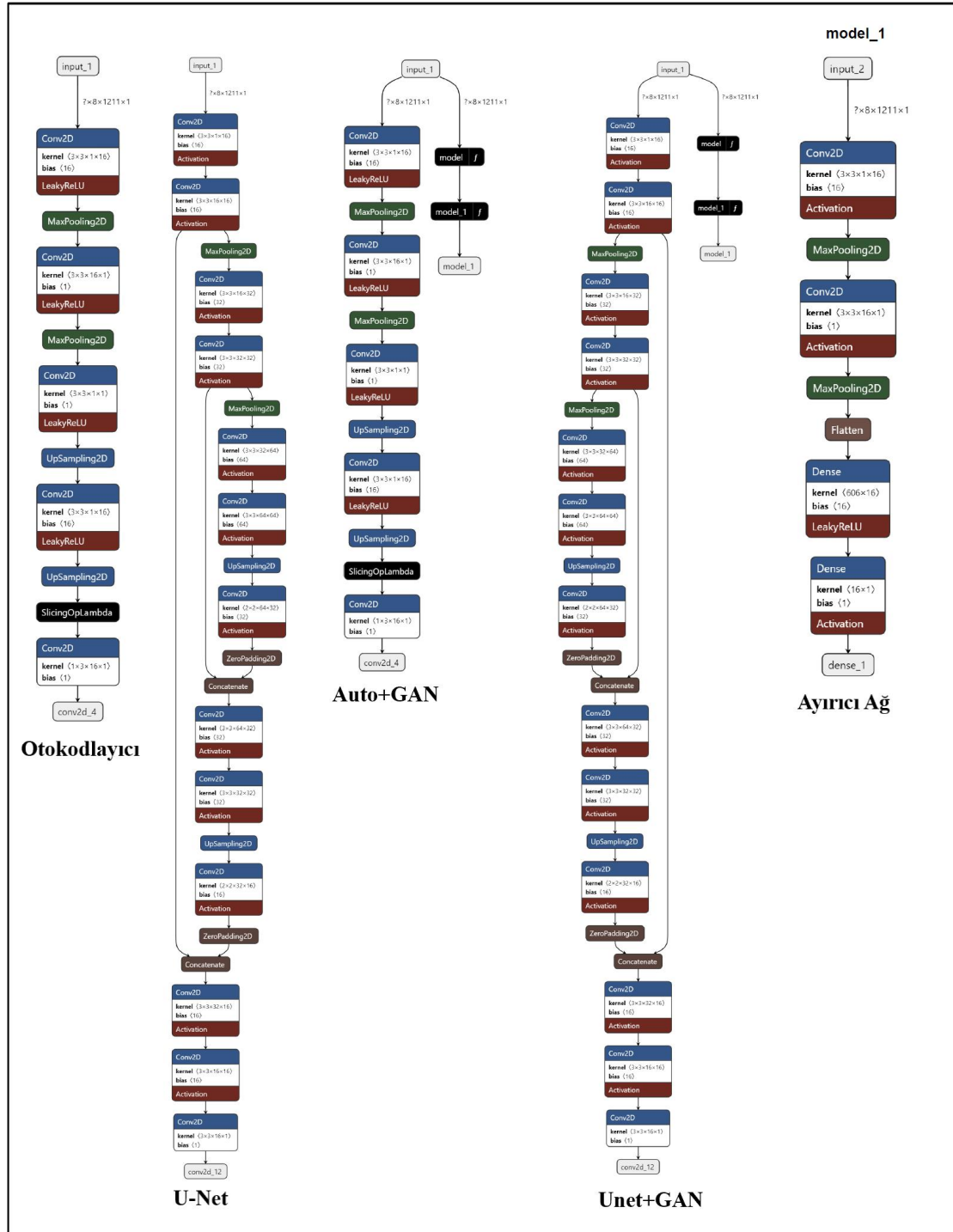
3.2.3. Çalışmada Kullanılan Denetimsiz Derin Öğrenme Mimarileri

Düşük örnekli mikrobiyom verilerinin sınıflandırma performansını iyileştirmek için üç farklı DL yönteminden (Otokodlayıcı, U-Net ve GAN) yararlanılmış ve görüntü verilerine

dönüştürülen mikrobiyom verileri için dört farklı model geliştirilmiştir (Şekil 3.6). İlk model, 2B-CNN mimarisi kullanılarak oluşturulmuş bir otokodlayıcıdır. Kodlayıcı bölümü, giriş verilerinin mekânsal boyutlarını azaltırken temel özellikleri yakalayan LeakyReLU aktivasyonları ve maksimum havuzlama işlemlerine sahip bir dizi evrişimli katmandan oluşmaktadır. Kod çözücü, kodlayıcıyla benzer katmanlara sahiptir fakat kodlayıcının katmanlarının ters yönde yansıması gibi düşünülebilir. Ayrıca kod çözücü, giriş verilerini yeniden oluşturmak için üst örnekleme fonksiyonunu kullanır, çıkış katmanı ise yeniden yapılandırılmış görüntüleri üretmek için doğrusal bir aktivasyon fonksiyonu kullanır.

İkinci model olan U-Net, kodlayıcı yolundan gelen yüksek çözünürlüklü özellikleri, kod çözücü yolundan gelen üst örnekleme çıktılarıyla birleştiren atlama bağlantılarına sahip bir kodlayıcı-kod çözücü yapısından oluşur. U-Net'in kodlayıcı kısmı, giriş verilerinin boyutunu kademeli olarak azaltırken, öznitelik haritalarının derinliğini artıran bir maksimum havuzlama işleminin takip ettiği 2B-CNN'leri kullanır. Kodlayıcı ve kod çözücü arasında bulunan U-Net'in darboğazı, derin öznitelikleri çıkarmaya odaklanan iki ek evrişimli katman içerirken herhangi bir havuzlama işlemi içermez. U-Net'in kod çözücü kısmı ise girişle eşleşebilmesi için boyutları kademeli olarak artıran, evrişimli katmanların takip ettiği üst örnekleme katmanlarını içerir. Kodlayıcıdan gelen öznitelik haritalarını birleştirmek için atlama bağlantıları uygulanır ve bu bağlantılar ağın kodlama işlemi sırasında kaybolabilecek mekânsal bilgileri korumasına olanak tanır. Son olarak, giriş görüntülerini yeniden yapılandırarak çıktıyı üretmek için doğrusal aktivasyon fonksiyonuna sahip bir evrişimli katman kullanılmıştır.

Üçüncü model, Auto+GAN modeli, bir ayırıcı ağ ile eşleştirilmiş bir otokodlayıcı içerir ve çekişmeli eğitim sunarak yeniden yapılandırma kalitesini iyileştirmek üzere tasarlanmıştır. Otokodlayıcı mimarisi daha önce açıklanan mimariye benzerken, ayırıcı mimarisi, ReLU aktivasyonlu 2B-CNN'lerden, maksimum havuzlama katmanlarından ve tam bağlı bir katmandan oluşmaktadır. Eğitim sırasında, otokodlayıcı yeniden yapılandırma kaybını (MSE) en aza indirmeyi hedeflerken, ayırıcı ağ ise gerçek ve yeniden yapılandırılmış görüntüleri doğru şekilde sınıflandırma yeteneğini en üst düzeye çıkarmaya çalışır. Auto+GAN modeli, otokodlayıcı kaybının %90 ve ayırıcının sınıflandırma kaybının %10 katkıda bulunduğu bir ağırlıklı toplam kayıp fonksiyonu



Şekil 3.6. Deeptaxim’de kullanılan derin öğrenme modellerinin detaylı katman bilgileri.

kullanılarak eğitilmiş, eğitim sırasında otokodlayıcının ağırlıkları, ayırcının eğitimi sırasında güncellenmeyi önlemek için dondurulmuştur.

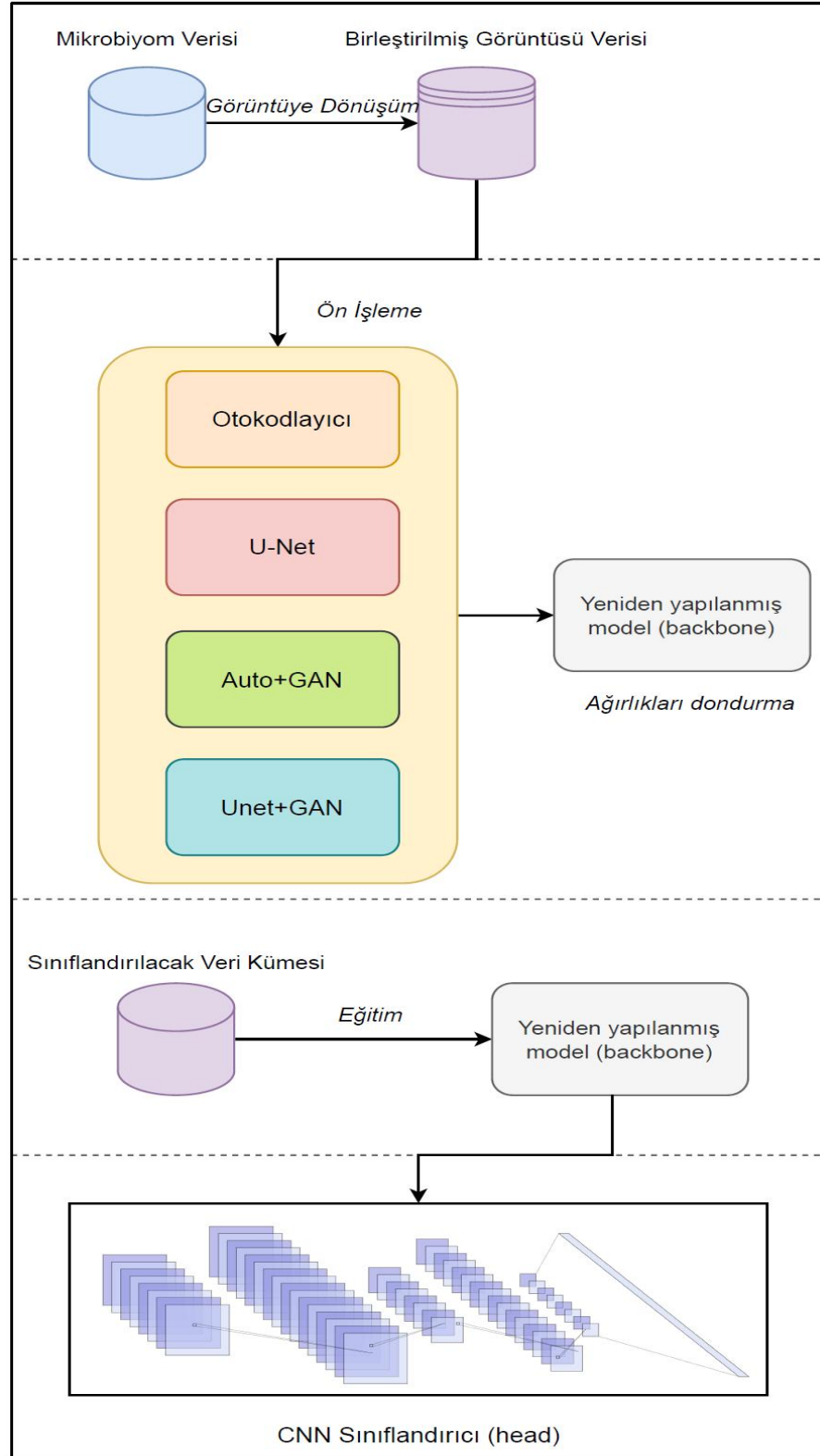
Dördüncü model, Unet+GAN, mikrobiyom görüntü verilerinin yeniden yapılandırılmasını iyileştirmek için bir GAN çerçevesi entegrasyonu U-Net mimarisi

üzerine inşa edilmiştir. U-Net, GAN'da giriş verilerini yeniden yapılandırmakla görevli üretici ağ olarak görev yapar. Auto+GAN modeline benzer şekilde yapılandırılmış olan ayırıcı ağ, gerçek ve yeniden yapılandırılmış görüntüler arasında ayrım yapmak üzere eğitilmiştir. U-Net'in atlama bağlantıları, özellikle ayırıcının çekişmeli kaybıyla eşleştirildiğinde yeniden yapılandırma süreci sırasında, mekânsal bilgileri korumada önemli bir rol oynamıştır. Model, U-Net'in yeniden yapılandırma kaybını GAN'ın çekişmeli kaybıyla dengeleyen bir kayıp fonksiyonu kullanılarak eğitilmiş, sınıflandırma kaybı için ağırlık 0,1'e, yeniden yapılandırma kaybı için ağırlık ise 0,9'a ayarlanmıştır.

U-Net modeli 0,001 öğrenme oranına sahip Adam optimizasyon algoritmasını kullanırken, kalan üç model için Adamax algoritması kullanılmıştır. Tüm modeller için eğitim, 16'lık bir parti boyutu ile 20 döngü boyunca gerçekleştirilmiş ve kayıp fonksiyonu olarak MSE kullanılmıştır. Bahsedilen tüm parametreler, eğitimden önce her model için ayrı bir parametre optimizasyon analizi gerçekleştirilerek belirlenmiş, dört modelin parametrelerini optimize etmek için tensorflow scikit-learn kütüphanesinden KerasRegressor sarmalayıcısı kullanılmıştır. Optimizasyon algoritmalarının (adam, adamax), aktivasyon fonksiyonlarının (ReLU, ELU ve LeakyReLU), kayıp fonksiyonlarının (MSE, ikili çapraz entropi ve MAE) çeşitli kombinasyonları ve ayrıca döngüler (20, 30, 50 ve 100) ve parti boyutları (8, 16, 32, 64 ve 128) için farklı değerleri keşfetmek için bir parametre ızgarası tanımlanmıştır. Bu yaklaşım, her bir özel görev için en uygun hiperparametre kümesini belirlemek üzere modellerin sistematik olarak ayarlanmasına olanak sağlamıştır.

3.2.4. Deeptaxim Modelinin Mikrobiyom-görüntü Veri Kümelerine Uygulanması

Uygulama, mikrobiyom verilerinin taksonomik sınıflandırmaya dayalı görüntü gösterimlerine dönüştürülmesiyle başlar. Bu görüntü verileri, Otokodlayıcı, U-Net, Auto+GAN veya Unet+GAN gibi DL modellerine dayalı Deeptaxim için girdi görevi görür. İlk olarak model, verilerde bulunan yapısal ve dahili kalıpları yakalamayı ve temsil etmeyi öğrendiği birleştirilmiş bir veri kümesi üzerinde denetimsiz bir şekilde eğitilir. Eğitilen modeller, alt akış görevleri için sınıflandırıcı başlıklarla (head) birleştirilmek üzere temsili öğrenme omurgası (backbone) olarak kullanılır (Şekil 3.7). Sürecin son aşamasında, eğitilen "Deeptaxim" modelinin performansını değerlendirmek için



Şekil 3.7. Deeptaxim sürecinin genel görünümü.

sınıflandırıcı model önemli bir rol oynar. Sınıflandırıcı modeli, ReLU aktivasyonunu kullanan (3x3) boyutunda 16 filtreye sahip, her biri öznetelik haritalarının mekânsal

boyutlarını azaltmak için bir maksimum havuzlama katmanı ile takip edilen, iki adet 2B-CNN ile oluşturulmuştur. Düzleştirilmiş çıktı daha sonra sigmoid aktivasyon fonksiyonuna sahip tam bağlı katmana beslenerek ikili bir sınıflandırma çıktısı üretir.

Sınıflandırıcı, Adamax optimizasyon algoritması ve kayıp fonksiyonu olarak MSE kullanılarak eğitilir. Sağlam bir performans sağlamak için, her katlamadaki sınıf oranını koruyan katmanlı 5 katlı çapraz doğrulama kullanılır. Eğitim sırasında aşırı uyumu önlemek için, erken durdurma uygulanır ve modelin performansı doğrulama kaybına göre izlenir. Performans değerlendirmesi üç temel ölçüm kullanılarak gerçekleştirilir: AUC, doğruluk (accuracy) ve F1 puanı. Tüm modeller için kapsamlı bir değerlendirme sağlamak adına, bu ölçümler her katlama için hesaplanır ve ortalama değerleri raporlanır.

3.2.5. Deeptaxim Yöntemiyle Elde Edilen Sınıflandırma Analizi Sonuçları

3.2.5.1. Gastro Veri Kümesinde Deeptaxim ve Mevcut Yöntemlerin Sınıflandırma Performansının Karşılaştırılması

Deeptaxim yönteminin GI hastalıklar üzerindeki sınıflandırma performansı, literatürdeki yöntemlerle, yani RF, XGB ve MLP ve iMic [164] yöntemleri ile karşılaştırılmıştır. Bu hesaplama için, birden fazla hastalık sınıfına sahip üç İBH veri kümesi, veri kümelerinde ikili sınıflandırma gerçekleştirmek için tek bir hastalık sınıfında birleştirilmiştir. İlk olarak, önerilen modelin eğitimi için, 2377 örnek ve 1104 öznitelik içeren birleştirilmiş veri kümesi %20 test verisi ve %80 eğitim verisi olarak bölünmüş, test verileri bu işlem sırasında ayrı tutularak, geliştirilen Otokodlayıcı (Auto), U-Net, Auto+GAN veya U-net+GAN modelleri, eğitim verilerine uygulanmıştır. Ardından, birleştirilmiş veri kümesiyle eğitilen modelin, İBH ve KK hastalıkları üzerindeki sınıflandırma performansı ölçülmüştür. Bu değerlendirme için, eş öznitelik, farklı örnek sayılarına sahip olan ve ilk veri setine benzer şekilde birleştirilen, İBH (1066 örnek) ve KK (360 örnek) veri kümeleri kullanılmıştır.

Güçlü bir sınıflandırma performansı değerlendirmesi sağlamak için, tabakalı 5 katlı çapraz doğrulama yöntemi kullanılmıştır. Tablo 3.3'te raporlanan tüm AUC, doğruluk (ACC) ve F1 puanları, bu beş çapraz doğrulama sonucunun ortalamasıdır. Karşılaştırma

için seçilen yöntemlerle ilgili ayrıntılar şu şekildedir: RF sınıflandırıcısı için, ormandaki ağaç sayısı 900 ve bir iç düğümü bölmek için gereken minimum örnek sayısı 2 olarak ayarlanır. XGB sınıflandırıcısı için, gerçekleştirilecek yükseltme adımı sayısı 900 olarak ayarlanır. Ayrıca, MLP sınıflandırıcısı için gizli katman boyutu 100, aktivasyon fonksiyonu ReLu ve maksimum yineleme sayısı 1000'dir. Son olarak iMic için, en verimli (ilgili orijinal çalışmada belirtildiği gibi) sınıflandırma sonucunu gösteren model (iki CNN katmanı ve ardından bir tam bağlı katmandan oluşan model) uygulanır.

Tablo 3.3. Deeptaxim' in GI hastalıklar üzerindeki sınıflandırma performansının mevcut yöntemlerle (RF, XGB, MLP ve iMic) karşılaştırılması.

		RF	XGB	MLP	iMic	Auto	U-Net	Auto+GAN	Unet+GAN
İBH	AUC	0.805	0.846	0.818	0.906	0.9	0.927	0.902	0.925
	ACC	0.865	0.881	0.895	0.856	0.869	0.878	0.863	0.875
	F1	0.908	0.915	0.854	0.897	0.905	0.913	0.903	0.912
KK	AUC	0.823	0.8	0.708	0.878	0.897	0.906	0.889	0.899
	ACC	0.822	0.8	0.708	0.8	0.819	0.839	0.839	0.836
	F1	0.812	0.794	0.701	0.798	0.816	0.838	0.837	0.835

Tablo 3.3'te raporlanan sonuçlara göre, İBH ve KK için, iMic ve Deeptaxim modelleri ile elde edilen sınıflandırma performansının alternatif modellere göre üstün olduğu görülmektedir. Başka bir deyişle, kladogram bilgilerinin CNN tabanlı çerçeveler ile görüntü verisi olarak gömülmesi, İBH ve KK veri kümeleri üzerindeki sınıflandırma performansını önemli ölçüde artırmıştır. Bununla birlikte, burada önerilen yöntem gibi karmaşık DL mimarileri daha fazla sayıda örnekle eğitildiğinde daha yüksek performans göstermesi muhtemeldir. Bu amaçla, sonraki deneyler Çoklu hastalık veri kümesiyle yürütülmüştür.

3.2.5.2. Çoklu Hastalık Veri Kümesinde Deeptaxim ve Mevcut Yöntemlerin Sınıflandırma Performansının Karşılaştırmalı Analizleri

3.2.5.2.1. Deeptaxim ile Sağlık-Hastalık (Wellness) İkili Sınıflandırma Analizi

Önerilen yöntem, mikrobiyom verileri üzerinde doğru ve genelleştirilebilir performans

Tablo 3.4. Deeptaxim' in Çoklu hastalık veri kümesi üzerindeki sınıflandırma (İS: İkili sınıflandırma (Sağlıklı/Hasta)) performansının mevcut yöntemlerle (RF, XGB, MLP, LR ve iMic) karşılaştırılması.

		RF	XGB	MLP	LR	iMic	Auto	U-Net	Auto+GAN	Unet+GAN
İS	AUC	0.754	0.799	0.803	0.804	0.879	0.878	0.888	0.883	0.891
	ACC	0.815	0.842	0.84	0.806	0.818	0.817	0.83	0.827	0.834
	F1	0.666	0.73	0.733	0.72	0.683	0.686	0.713	0.709	0.713

değerlendirmesi için 11 farklı hastalıktan 8069 örnek içeren daha kapsamlı bir veri kümesinde test edilmiş, RF, XGB, MLP, iMic ve GMWI2 [165] yöntemleriyle karşılaştırılmıştır. İlk olarak, görüntü verisine dönüştürülen birleştirilmiş çoklu hastalık veri seti, %20 test verisi ve %80 eğitim verisi olarak ikiye bölünerek eğitim verilerine otokodlayıcı (Auto), U-Net, Auto+GAN veya Unet+GAN modelleri uygulanmış, daha sonra birleştirilmiş veri setiyle eğitilen modelin hastalıklı örnekler ve sağlıklı kontrol grupları üzerinde sınıflandırma performansı ölçülmüştür. Mevcut yöntemlerin parametreleri bir önceki bölümdekiyle aynıdır. GMWI2 için, orijinal çalışmada belirtildiği gibi parametreleri şu şekilde raporlanan Doğrusal Regresyon (LR) yöntemi kullanılmıştır: $L1$ (Lasso) cezası Liblinear çözücü ile kullanılmış, rastgele durum 42 olarak ayarlanmıştır, düzenleme parametresi C , 0,03 olarak ayarlanmış ve sınıf ağırlığı "dengeli" olarak seçilmiştir. Tablo 3.4'te gösterilen her AUC, doğruluk (ACC) ve F1 puanı, tabakalı 5 katlı çapraz doğrulama sonucunun ortalamalarıdır. Elde edilen sonuçlara göre RF, XGB, MLP ve LR yöntemlerinin sınıflandırma performanslarının CNN tabanlı yöntemlere göre düşük olduğu ve önerilen yöntemin, özellikle UNet+GAN modelinin, diğer yöntemlere göre daha üstün bir performans sergilediği söylenebilir.

3.2.5.2.2. İkili Sınıflandırmanın Hastalığa Özgü Fenotip Analizi

Modellerin her hastalık kategorisini (Aterosklerotik kardiyovasküler hastalık (ACVD), Ankilozan spondilit (AS), Kolorektal kanser (KK), Crohn hastalığı (CH), Graves hastalığı (GH), Karaciğer Sirozu (KS), Multipl skleroz (MS), Non-Alkolik Yağlı Karaciğer Hastalığı (NAYKH), Romatoid Artrit (RA), Tip 2 diyabet (T2D) ve Ülseratif kolit (ÜK)) tahmin edebilme kapasitesini ölçmek amacıyla bir değerlendirme analizi

gerçekleştirilmiştir. Bu değerlendirme için, önceki bölümde anlatılan ikili sağlık analizi için kullanılan modeller kullanılmış ve her alt sınıf (hastalık) için sınıflandırma performansı gözlemlenmiştir.

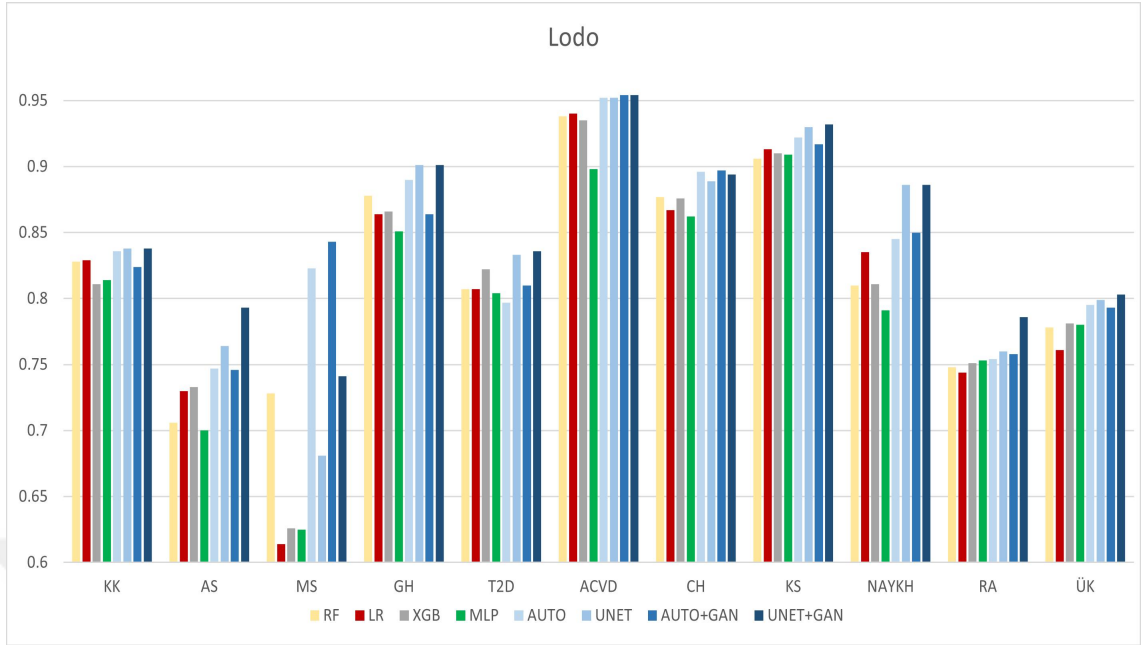
Tablo 3.5. Sağlık değerlendirmesinin her model için hastalık başına doğru tahmin yüzdesi.

	Auto	U-Net	Auto+ GAN	Unet+ GAN	RF	XGB	MLP	LR
Sağlıklı	90.01	89.94	90.19	90.9	89.7	89.84	88.16	81.09
ACVD	86.92	86.45	85.98	85.98	85.05	85.92	84.11	85.33
AS	37.89	45.26	33.68	46.32	21.05	44.21	44.84	56.32
KK	74.14	75.67	73.26	75.41	66.03	72.37	74.81	75.17
CH	69.72	72.54	70.77	71.13	69.72	68.87	70.77	81.34
GH	64	72	69	75	78	77	78	77
KS	77.63	84.87	78.29	80.92	78.58	78.89	78.87	78.11
MS	41.67	29.17	37.5	25	25	41.67	20.83	50
NAYKH	52.33	55.81	53.49	53.49	48.84	52.33	53.95	56.23
RA	37.09	43.05	35.76	41.06	31.13	33.74	43.64	46
T2D	54.38	51.72	50.93	53.05	45.36	49.95	52.42	62.15
ÜK	49.6	58	46.4	52.4	42.8	50.4	52	54.8

Tablo 3.5’te bu gözlemleri içeren analiz sonuçları verilmiştir. Önerilen yöntem, özellikle sağlıklı kontroller de dahil olmak üzere önemli örneklem büyüklüklerine sahip hastalıkları ayırt etmede diğer modellere kıyasla daha iyi performans göstermiştir.

3.2.5.2.3. Bir Hastalığı Dışarıda Bırakma Yaklaşımının Değerlendirilmesi

Bu bölümde, sınıflandırma görevi için bir hastalığı dışarıda bırakma (Leave-one-disease-out, Lodo) yaklaşımı kullanılmıştır. Öncelikli amaç, Deeptaxim’in farklı hastalıklar için sıfır-atış öğrenme (zero-shot learning) performansını değerlendirmektir. Bu amaçla, her bir deneyde, on bir hastalıktan biri ve ayrıca bu hastalığın örnek sayısının iki katı oranda rastgele seçilmiş sağlıklı örnek verisi eğitimden hariç tutulmuştur. Modeller, geri kalan on hastalık ve sağlıklı kontrollerden örnekler



Şekil 3.8. Çoklu hastalık veri kümesinde, bir hastalığı dışarıda bırakma (Lodo) yaklaşımının değerlendirilmesi.

içeren bir veri kümesi üzerinde eğitilmiş ve dışarıda bırakılan hastalık ve bir dizi sağlıklı kontrol örneklerini içeren bir veri kümesi üzerinde test edilmiştir. Tüm modeller için AUC değerlendirme sonuçları Şekil 3.8’de verilmiştir. Sonuçlar, Lodo yaklaşımında, Deeptaxim modelinin tüm hastalıklar için diğer modellere göre üstünlüğünü ve çeşitli hastalık bağlamlarında genelleştirilme potansiyelini vurgulamaktadır.

4. BÖLÜM

MİKROBİYOM VERİLERİNDE BİYOBELİRTEÇ KEŞFİ

4.1. Disrobiom: Mikrobiyom Verileri için Derin Öğrenme Algoritmalarıyla Gürbüz Biyobelirteç Keşfi

4.1.1. Genel Bakış

Biyolojik verilerden elde edilen sonuçlarda tekrarlanabilirliğin sağlanması, klinik uygulamalar ve yanlış biyolojik bulguların önlenmesi için büyük önem taşır, ancak boyutluluğun laneti, bağırsak ekosistemlerindeki önemli farklılıklar, metagenomik verilerdeki dizileme hataları gibi çeşitli zorluklar nedeniyle hastalık tespiti için gürbüz biyobelirteçler bulmak zorlaşmaktadır. Bu nedenle, önerilen çalışma, İBH ve KK hastalıkları için az sayıda örneğe rağmen yüksek ve tutarlı sınıflandırma performansı gösteren bir model geliştirmeyi ve tanı veya tedavi işlemlerinde kullanılabilecek potansiyel kararlı/gürbüz biyobelirteçleri tanımlamayı amaçlamaktadır. Bu amaçla geliştirilen Disrobiom modeli, yüksek boyutlu, düşük örnekli bağırsak mikrobiyomu veri kümelerinde maksimum verimlilikle sınıflandırma yapmak için denetimsiz DL modellerini kullanır. Doğrusal modellere kıyasla daha yüksek karmaşıklığa sahip otokodlayıcı ve U-Net modellerinin, metagenomik girdi verilerinin temsili verimli bir şekilde öğrenebildiği, bunu olabildiğince sadık bir şekilde yeniden yapılandırabildiği ve giriş ve çıkış verileri arasındaki bu küçük farktan gürbüz biyobelirteçleri tespit edebildiği gözlemlenmiştir.

4.1.2. Kullanılan Veri Kümeleri ve Ön İşlemler

Çalışmada, ENA moleküler veri tabanı tarafından sağlanan altı adet kamuya açık insan

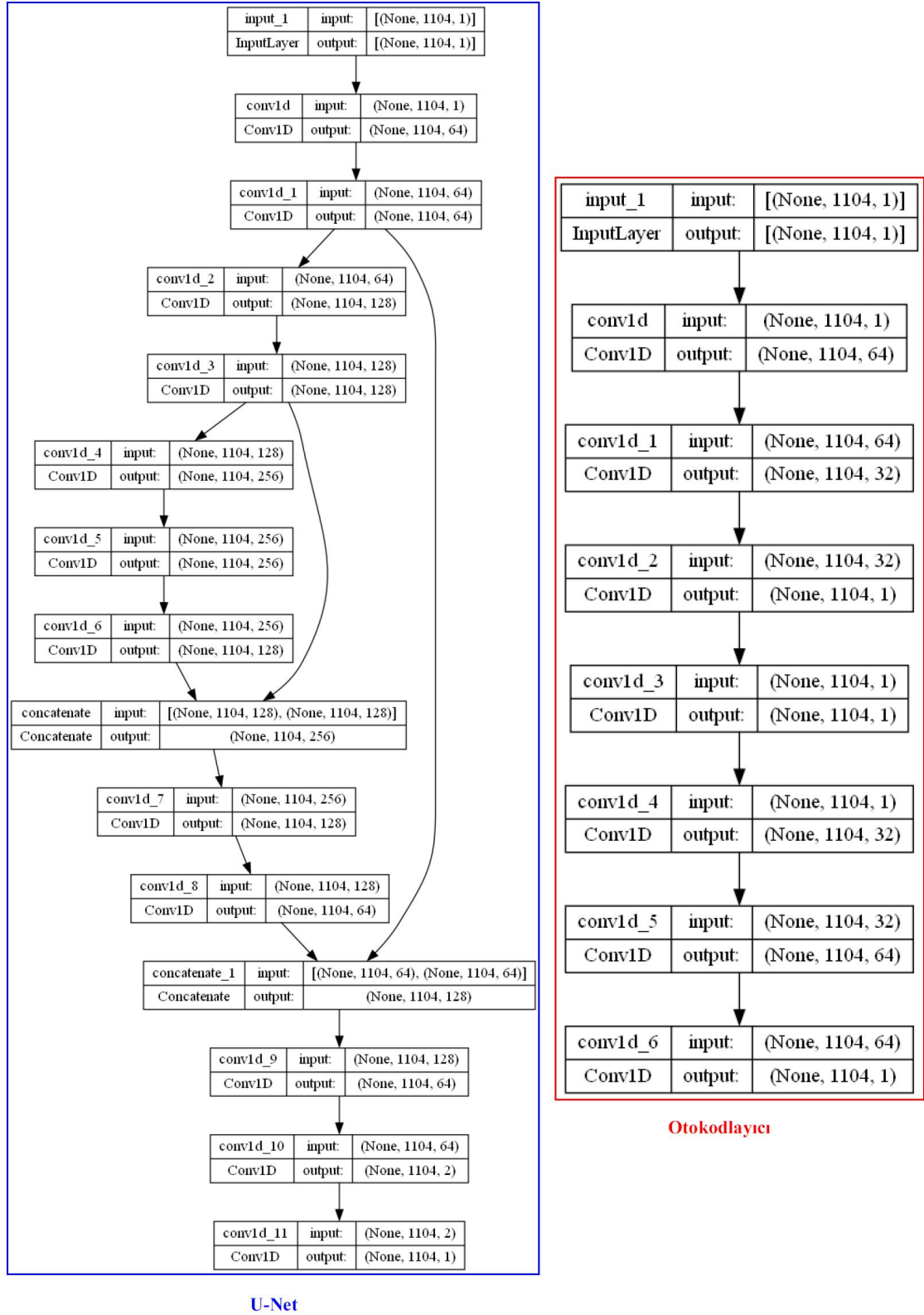
bağırsağı metagenomik veri seti kullanılmıştır. Çeşitli araştırma makalelerinden toplanan bu altı veri kümesinden üçü İBH [155–157] kalan üç veri seti ise KK kohortlarından oluşmaktadır [159–161]. Her bir veri seti için hastalık örneklerinin sayısı, sağlıklı kontrollerin sayısı, toplam örnek sayısı ve veri setlerine verilen adlar Tablo 4.1’de verilmiştir. Elde edilen veri kümelerini işleyerek taksonomik öznitelikleri çıkarmak için, MetaPhlAn 3.0 [163] kullanılmıştır. Veri setlerindeki özniteliklerin çoğu ortak olmakla birlikte, tüm veri setleri için ortak olmayan (çoğunluk için ortak) birkaç öznitelik de eklenerek her veri seti için toplamda 1104 ortak taksonomik öznitelik elde edilmiştir. Tablo gösteriminde, CNN modellerinin kısa vadeli korelasyonları yakalamasına yardımcı olmak için, taksonomik birimler yakınlardaki benzer taksonları mekânsal olarak gruplamak üzere isim sırasına göre sıralanmıştır.

Tablo 4.1. Veri Kümelerine ait detaylı bilgiler.

Veri Kümesi	Toplam Örnek Sayısı	Hasta Örnekleri	Sağlıklı Kontrol Sayısı	Çalışma Ref.
İBH-I	213	ÜK=72, CH=85	56	[155]
İBH-II	314	ÜK=88, CH=170	56	[156]
İBH-III	539	ÜK=242, CH=69	228	[157]
KK-I	112	46	66	[159]
KK-II	130	70	60	[160]
KK-III	118	53	53	[161]

4.1.3. Denetimsiz Derin Modeller

İBH ve KK için gürbüz biyobelirteçleri keşfetmek amacıyla iki farklı denetimsiz DL yöntemi (Otokodlayıcı ve U-Net) kullanılmıştır. X , otokodlayıcının giriş verisi ve X' çıkış verisi olarak kabul edilirse, otokodlayıcı ağının temel eğitim amacı, X ve X' arasındaki yeniden yapılandırma hatasını en aza indirmektir. Matematiksel olarak, modelin çıktısı $X' = g[f(X)]$ şeklinde ifade edilebilir. Burada $f(\cdot)$: kodlayıcı, $g(\cdot)$ kod çözücü fonksiyonudur ve amaç fonksiyonu $L(X, X') = \|X - X'\| = \|X - g[f(X)]'\|$ şeklinde



Şekil 4.1. Derin modellerin detaylı katman bilgileri.

tanımlanabilir. Temel olarak hem otokodlayıcı hem de U-Net modelleri giriş verileriyle eğitilir (çalışmamızda, metagenomik veri kümeleri) ve giriş verilerini minimum kayıpla

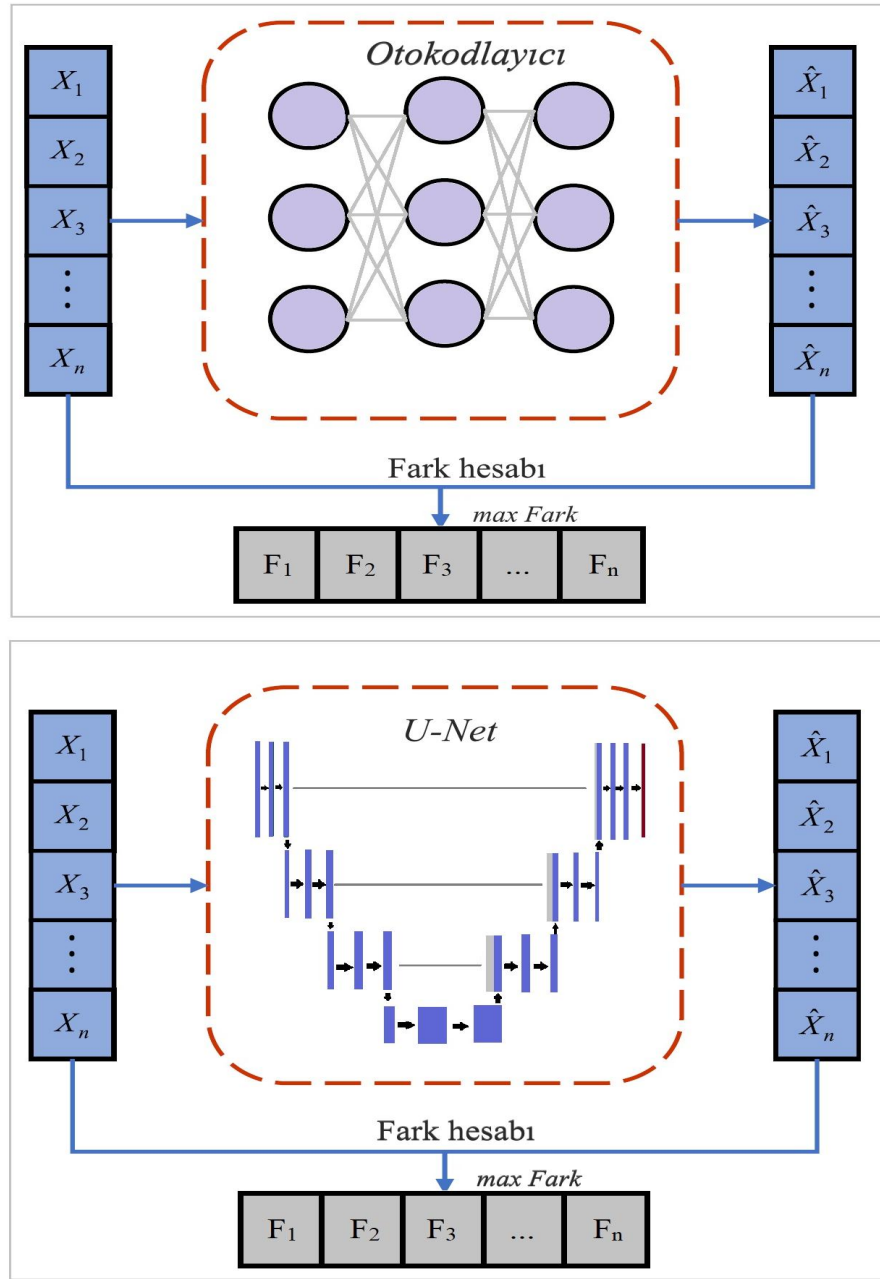
gizli uzaydan yeniden oluşturmaya çalışır. Ancak bu şekilde elde edilen çıktı verileri, giriş verilerinden bir miktar farklı olacaktır. Bu çalışmada, modellerin girdi ve çıktıları arasında en büyük farka sahip değerlerin potansiyel biyobelirteçleri bulmada yararlı olabileceği öne sürülmektedir.

Her biri farklı sayıda örneğe ancak eşit sayıda öznitelige sahip altı farklı veri kümesi (İBH-I, İBH-II, İBH-III, KK-I, KK-II, KK-III), otokodlayıcı ve U-Net modellerini eğitmek için kullanılmıştır. Otokodlayıcı ve U-Net mimarileri için tam bağlı katmanlar yerine, hem kodlayıcı hem de kod çözücüde aynı sayıda tek boyutlu Evrişimli Sinir Ağları (1B-CNN) kullanılmıştır. U-Net ve otokodlayıcı modellerinin detaylı katman gösterimleri Şekil 4.1'de gösterilmiştir. Her iki model için de aktivasyon fonksiyonu, optimizasyon algoritması ve modellerde kullanılacak parti boyutu gibi parametreleri belirlemek için bir ön çalışma yürütülmüş ve hiperparametre optimizasyonu sağlanmıştır. Bu çalışmanın sonucunda, belirlenen değer arasından (örneğin, L_1 katsayısı için [0.1, 0.01, 0.001, 0.0001], parti boyutu için [16, 32, 64, 128], optimizasyon algoritması için Adam ve Adamax) Otokodlayıcı ve U-Net modelleri için en iyi performansı sağlayan hiper parametreler, model parametreleri olarak seçilmiştir. Sigmoid aktivasyon fonksiyonu yalnızca çıkış katmanında kullanılırken, tüm 1B-CNN katmanları için aktivasyon fonksiyonu olarak LeakyReLU kullanılmıştır. Ek olarak, optimizasyon fonksiyonu olarak Adamax ve kayıp fonksiyonu olarak özel bir kayıp fonksiyonu (L_1 katsayısı 0.0001 olan MSE) kullanılmıştır. Son olarak, parti boyutu 32 ve döngü sayısı 50 olarak ayarlanmıştır. Parametrelerde olduğu gibi, gizli uzay boyutu için de bir ön çalışma yapılmış ve çeşitli varyasyonlar (B, B/2, B/4 ve B/8, boyut B=1104 için) arasında gizli uzay boyutu, giriş verileriyle aynı olacak şekilde (B) seçilmiştir.

4.1.4. Derin Modellerle Öznitelik Seçimi

Öznitelik Seçimi (FS) sürecinden önce, veri kümeleri her iki modeli de eğitmek için 8:2 oranında eğitim ve test kümelerine ayrılır. Modeller eğitildikten sonra, ortaya çıkan model çıktısı X' ve giriş verisi X bir matris olarak kabul edilir ve aralarındaki fark (hata matrisi, HM) şu şekilde hesaplanır: $HM = ||X - X'||$.

Daha sonra, HM matrisindeki en büyük değerlerin indeksleri (en yüksek farkı oluşturan

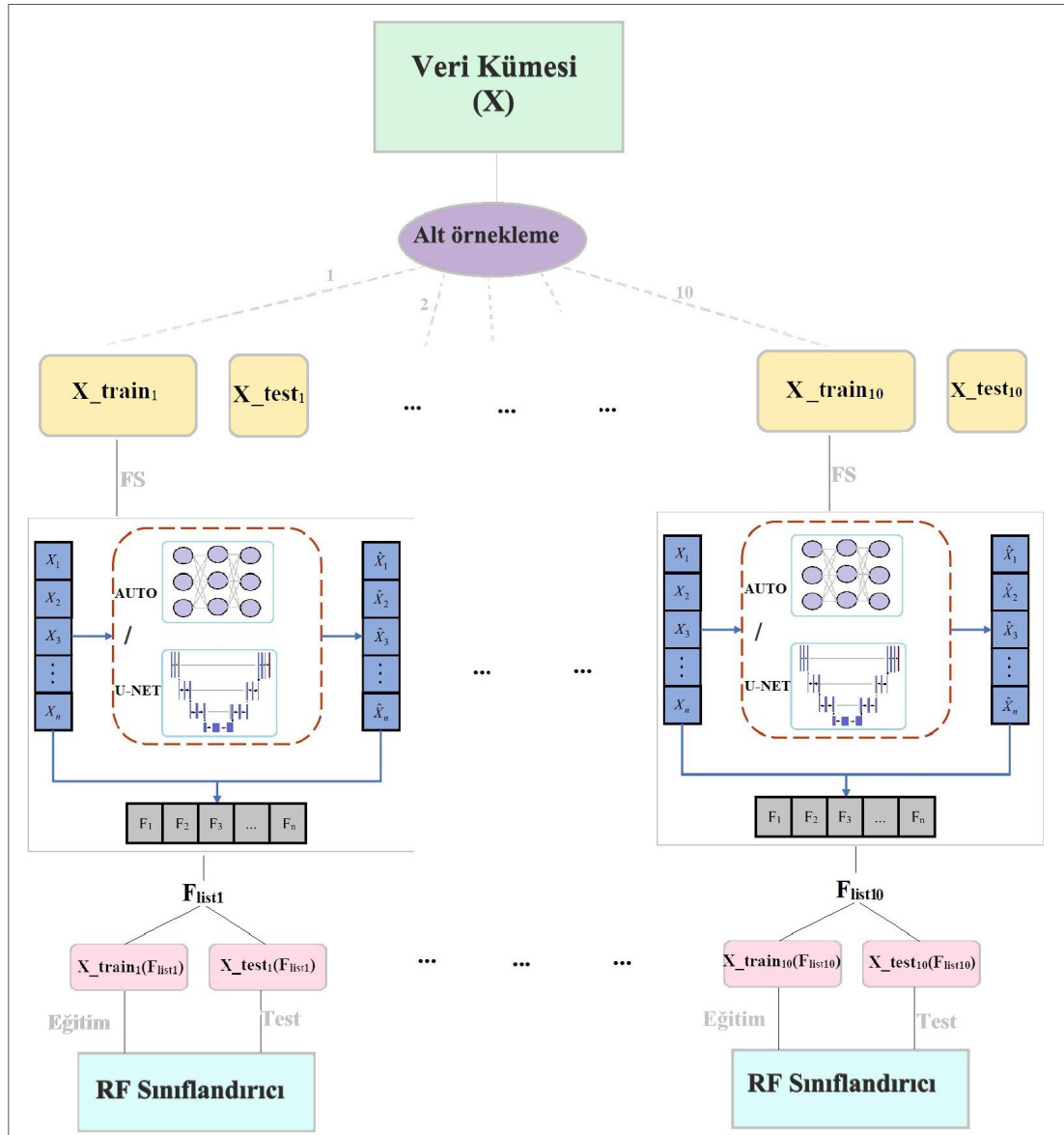


Şekil 4.2. Modellerin girdi ve çıktıları arasındaki fark kullanılarak öznelilik seçme prosedürü.

öznelilikler) farkların büyüklüğüne göre sıralanır. HM, tıpkı X ve X' gibi, eğitim veri kümesiyle aynı boyutlara sahiptir: [örneklem boyutunun %80'i, 1104 (öznelilik boyutu)]. Bu nedenle, örneğin, HM 'deki en yüksek değerlere sahip öznelilik indeksleri sıralı listesinin ilk 10 ögesi ($F_n = F_{10}$ ise) bize 10 potansiyel biyobelirteci sağlayacaktır. Otokodlayıcı ve U-Net modelleri için FS süreci Şekil 4.2'de gösterilmiştir.

Önerilen çerçevede [1]'e benzer bir performans değerlendirme protokolü izlenmiştir.

Temel motivasyon, orijinal veri kümelerinin farklı varyantlarında kararlılık ve sınıflandırma performansını ölçmeye dayanmaktadır. Modellerin sınıflandırma performansı ve tutarlılığı, gürbüz bir biyobelirteç tespit algoritmasının, veri kümesindeki popülasyon değişiklikleri altında benzer sonuçlar üretmesi gerektiği fikrine dayanarak değerlendirilmiştir. Bu şekilde, popülasyon alt örnekleme ihtiyaçlarına uygun farklı kohortlar oluşturmak, algoritma tarafından tanımlanan biyobelirteçler için temelde önemli değişikliklere neden olmayacaktır. Bu nedenle, modellerin sınıflandırma performansını adil bir şekilde değerlendirmek ve keşfedilen biyobelirteçlerin kararlılığını test etmek



Şekil 4.3. Derin modellerin Öznitelik Seçimi ve Sınıflandırma-Tutarlılık performans değerlendirme çerçevesi.

için, her veri kümesi 10 kez eğitim ve test kümelerine rastgele bölünmüş ve her model için 8:2 oranı korunmuştur. Daha sonra, Şekil 4.3'te gösterildiği gibi, her eğitim kümesi Otokodlayıcı veya U-Net modellerinden biri kullanılarak FS sürecini izlemiş ve her biri n öznitelik içeren toplam 10 öznitelik listesi ($Flist$) elde edilmiştir. Daha sonra her $Flist$, RF sınıflandırıcısı ile sınıflandırılmış ve bu 10 sonucun ortalaması hesaplanmıştır.

Öznitelik listeleri ($Flist$) arasındaki tutarlılığı ölçmek için Kuncheva indeksi (KI) [131] adı verilen bir benzerlik indeksi kullanılmıştır. $n = |Flist_i| = |Flist_j|$ (çalışmamız için $n=10, 20, 50$ veya 100) ve p orijinal veri setindeki öznitelik sayısı ($p=1104$) olmak üzere, iki listenin KI değeri aşağıdaki gibi hesaplanır:

$$KI(Flist_i, Flist_j) = \frac{(p \cdot |Flist_i \cap Flist_j| - n^2)}{n(p - n)} \quad (4.1)$$

KI değerleri, -1 ile 1 değer aralığı arasında olabilir. 1'e yakın değerler, iki listenin çok sayıda ortak öge içerdiğini gösterirken, negatif değerler biyobelirteçlerin rastgele seçildiğini gösterir. Tüm listeler ($m = 10$) için, çiftler arası $m(m - 1)/2$ adet benzerlik hesaplandıktan sonra, algoritmanın genel tutarlılığı (C_{avg}), tüm çiftler arası benzerliklerin ortalaması alınarak aşağıdaki gibi hesaplanabilir:

$$C_{avg} = \frac{2 \sum_{i=1}^m \sum_{j=i+1}^m KI(Flist_i, Flist_j)}{m(m - 1)} \quad (4.2)$$

4.1.5. Disrobiom Yöntemiyle Elde Edilen Değerlendirme Sonuçları

4.1.5.1. Disrobiom Sınıflandırma ve Tutarlılık Performansının Mevcut Yöntemlerle Karşılaştırılması

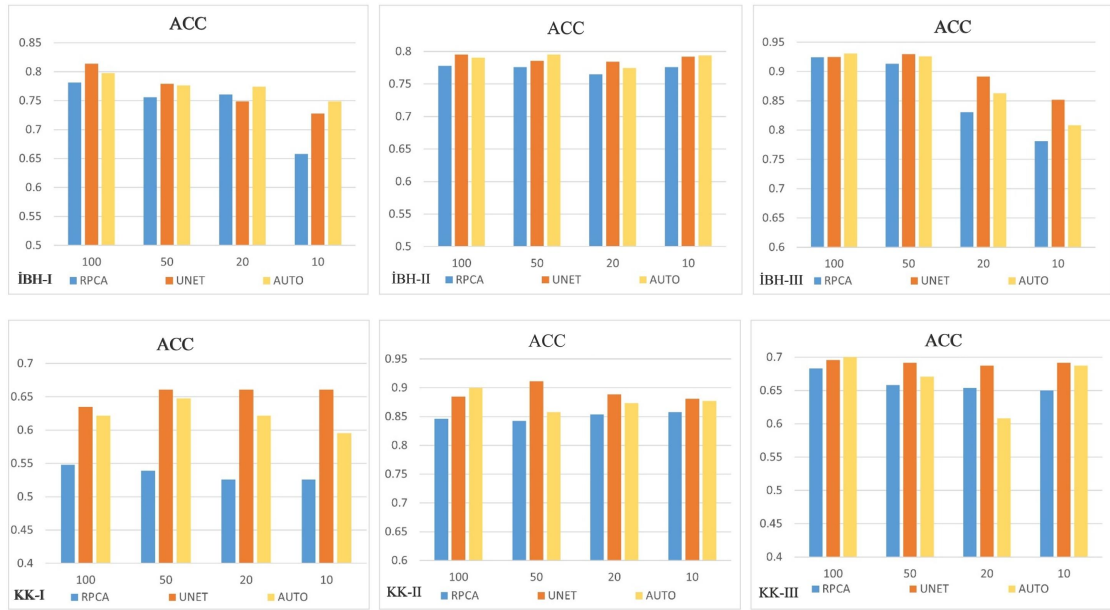
Önerilen yöntemin sınıflandırma ve tutarlılık performansı, yaygın olarak kullanılan yöntemlerle karşılaştırılmıştır. Karşılaştırma sonuçları, hastalık ve kontrol örnek grupları arasında ayırım yapmak için ikili sınıflandırma sonuçlarını göstermektedir. Bu inceleme için, birden fazla hastalık sınıfına (ÜK ve CH) sahip üç İBH veri kümesi, tüm veri kümelerinde ikili sınıflandırma yapabilmek adına tek bir hastalık sınıfında birleştirilmiştir. Farklı yöntemlerle keşfedilen 10, 20, 50 ve 100 özniteliğin performansını değerlendirmek

Tablo 4.2. Metagenomik veri kümeleri üzerinde RPCA, U-Net ve AUTO modelleri ile seçilen 10, 20, 50 ve 100 öznitelik için karşılaştırmalı RF Sınıflandırma Performansı (AUC ve F1 puanı).

	AUC			F1 puanı		
	RPCA	AUTO	UNET	RPCA	AUTO	UNET
İBH-I *100	0.7063	0.7307	0.75	0.8522	0.8605	0.8711
*50	0.6731	0.7038	0.7083	0.8368	0.8478	0.8508
*20	0.6807	0.7096	0.6689	0.8392	0.8435	0.8314
*10	0.5635	0.7007	0.6547	0.7746	0.8168	0.8129
İBH-II *100	0.5625	0.5844	0.5969	0.8704	0.8761	0.8794
*50	0.5614	0.5969	0.5781	0.8693	0.8793	0.8745
*20	0.5478	0.5563	0.575	0.8627	0.8688	0.8736
*10	0.5965	0.6144	0.6112	0.8651	0.8762	0.8754
İBH-III *100	0.9263	0.9336	0.9293	0.9345	0.9399	0.9347
*50	0.9141	0.9297	0.9261	0.9252	0.9357	0.9409
*20	0.83	0.8631	0.8944	0.8537	0.8815	0.9059
*10	0.7783	0.8021	0.8523	0.8113	0.839	0.8718
KK-I *100	0.5454	0.6021	0.615	0.4535	0.4961	0.5111
*50	0.5358	0.6325	0.6496	0.4397	0.5439	0.5572
*20	0.5404	0.6225	0.6525	0.4649	0.5354	0.5624
*10	0.5492	0.5967	0.6438	0.4786	0.5076	0.5461
KK-II *100	0.8039	0.8817	0.8595	0.8889	0.925	0.9143
*50	0.7983	0.8206	0.8984	0.8865	0.8965	0.9331
*20	0.815	0.8373	0.865	0.8941	0.9044	0.9171
*10	0.8388	0.8536	0.8539	0.8917	0.9077	0.9117
KK-III *100	0.6489	0.6867	0.6811	0.7561	0.7535	0.7521
*50	0.6089	0.67	0.6822	0.7465	0.7156	0.7433
*20	0.6189	0.6022	0.6878	0.7291	0.6658	0.731
*10	0.6533	0.6833	0.7022	0.6929	0.7365	0.7282

için altı farklı veri kümesi (İBH-I, İBH-II, İBH-III, KK-I, KK-II, KK-III) ve sınıflandırma performansını ölçmek için, ormandaki ağaç sayısı 900 ve bir iç düğümü bölmek için gereken minimum örnek sayısı 2 olan RF sınıflandırıcı kullanılmıştır.

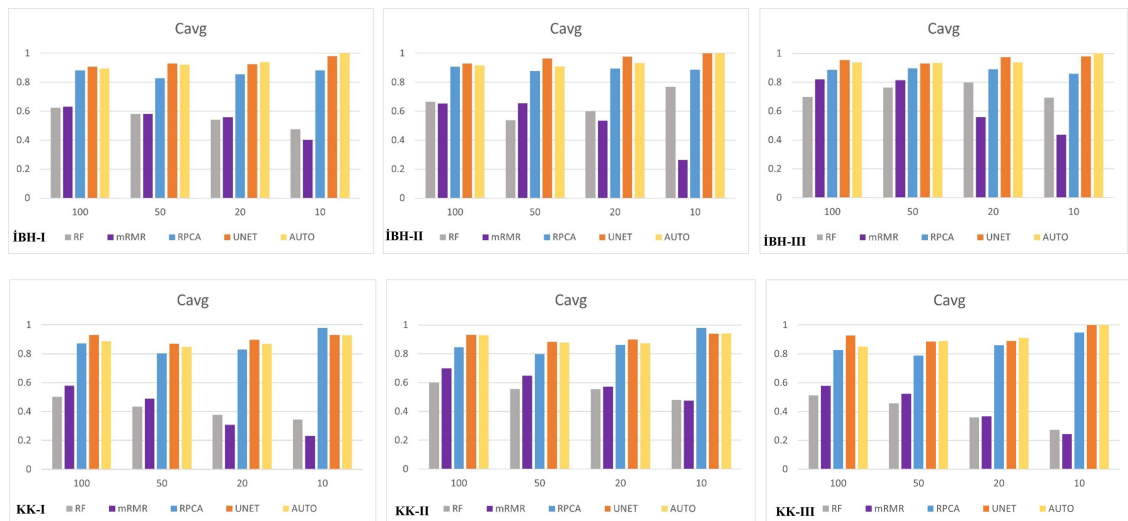
Sınıflandırma sonuçları U-Net ve Otokodlayıcı gibi denetimsiz bir yöntem olan RPCA [1] ile karşılaştırılmış ve Şekil 4.4'te görüldüğü gibi, İBH-I ve KK-III veri kümelerindeki 20 öznitelik koşulu haricinde hem Otokodlayıcı hem de U-Net ile seçilen öznitelikler, RPCA



Şekil 4.4. Metagenomik veri kümeleri üzerinde RPCA, U-Net ve Otokodlayıcı (AUTO) modelleri ile seçilen 10, 20, 50 ve 100 öznelik için karşılaştırmalı RF Sınıflandırma Performansı (ACC).

tarafından keşfedilen özneliklerden daha üstün sınıflandırma performansı göstermiştir. Şekil 4.4'te gösterilen her değer, veri kümelerinin on kez alt örneklemeyle elde edilen on deneyin ortalama doğruluk (ACC) puanını temsil eder. Sınıflandırma için daha ayrıntılı sonuçlar (AUC ve F1 puanı) Tablo 4.2'de verilmiştir.

FS için sıklıkla kullanılan denetimli yöntemlerden RF ve mRMR, önerilen yöntemin



Şekil 4.5. Metagenomik veri kümelerinde FS yöntemleri için tutarlılık performansı karşılaştırması.

tutarlılık performansını karşılaştırmak için eklenmiştir. Tüm yöntemler için, önerilen yöntem için açıklanan Cavg hesaplaması aynı şekilde gerçekleştirilmiştir. Şekil 4.5'te görülebileceği gibi, önerilen yöntem, KK-I ve KK-II veri kümelerindeki 10 öznitelik hariç tüm koşullar için diğer tüm yöntemlerden daha yüksek tutarlılık performansı göstermiştir.

RF ve mRMR yöntemleri ile 10 alt örneklemin her birinde seçilen öznitelik listeleri önemli ölçüde farklı olsa da (neredeyse tamamen farklı), her bir öznitelik listesi sınıflandırmada ayrı ayrı başarılı performans sergileyebilir. Bunun nedeni, bu yöntemlerin denetimli yöntemler olması ve evrensel belirteçleri keşfetmekten ziyade, ayırt edici değişkenleri doğru bir şekilde seçmeye daha fazla odaklanması ve bunun sonucunda ilişkili öğelerin tutarsız bir şekilde seçilmesiyle açıklanabilir. Bu nedenle, bu yöntemler karşılaştırmalı sınıflandırma performansı değerlendirmesinde denetimsiz modellerden ayrı tutulmuştur. Denetimli yöntemler, sınırlı özniteliğe sahip veri kümelerini başarılı bir şekilde sınıflandırma yeteneğine sahip olsa da gürbüz biyobelirteç keşfinde kullanılmak için uygun olmadıkları söylenebilir.

4.1.5.2. Keşfedilen Potansiyel Biyobelirteçler için Çapraz Kohort Doğrulaması

Bu bölümde, İBH-I, İBH-II, İBH-III, KK-I, KK-II, KK-III veri kümelerinden biri için Otokodlayıcı, U-Net ve RPCA yöntemleriyle keşfedilen en önemli 10 özniteliğin sınıflandırma performansı diğer beş veri kümesinde test edilmiştir. Şekil 4.6'da,

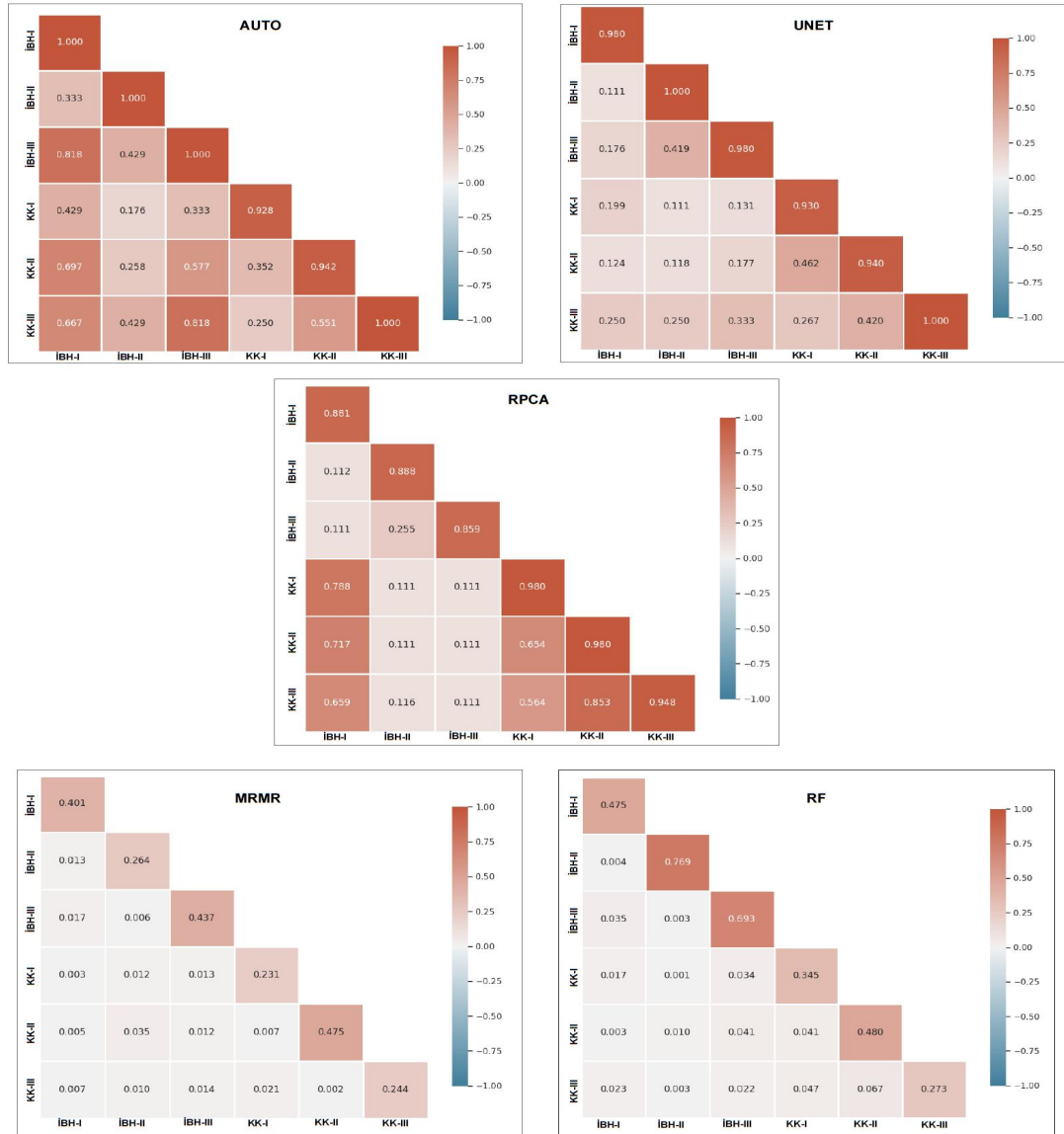
AUTO	İBH-I *	İBH-II *	İBH-III *	KK-I *	KK-II *	KK-III *
İBH-I	0.7488	0.7256	0.7233	0.7302	0.6558	0.7116
İBH-II	0.7962	0.7937	0.773	0.8079	0.7825	0.7794
İBH-III	0.7941	0.812	0.8083	0.8398	0.7572	0.7407
KK-I	0.6043	0.6087	0.6174	0.5957	0.6	0.5913
KK-II	0.8692	0.8577	0.8769	0.8577	0.8769	0.8423
KK-III	0.5167	0.5792	0.5583	0.6083	0.4792	0.6875

UNET	İBH-I *	İBH-II *	İBH-III *	KK-I *	KK-II *	KK-III *
İBH-I	0.7279	0.7279	0.7233	0.7047	0.6907	0.7023
İBH-II	0.7857	0.7921	0.8127	0.8175	0.8048	0.8143
İBH-III	0.8278	0.812	0.8519	0.8343	0.8389	0.8407
KK-I	0.6391	0.5739	0.587	0.6609	0.613	0.587
KK-II	0.8769	0.8154	0.8385	0.8808	0.8808	0.85
KK-III	0.6	0.5917	0.6042	0.6125	0.5417	0.6917

RPCA	İBH-I *	İBH-II *	İBH-III *	KK-I *	KK-II *	KK-III *
İBH-I	0.6581	0.7	0.6419	0.6791	0.7186	0.6977
İBH-II	0.7937	0.7762	0.8317	0.8	0.7635	0.781
İBH-III	0.8315	0.7907	0.7815	0.8333	0.7657	0.7241
KK-I	0.5913	0.5913	0.5522	0.5261	0.6	0.5826
KK-II	0.8692	0.8592	0.8308	0.8515	0.8577	0.85
KK-III	0.5333	0.5083	0.5792	0.5333	0.5375	0.65

Şekil 4.6. Bir veri kümesi için en önemli 10 özniteliğin, diğer veri kümeleri üzerindeki Otokodlayıcı, U-Net veya RPCA modelleri ile sınıflandırma sonuçları.

veri kümelerinden birinde (sıra indeksiyle gösterilen) denetimsiz modeller tarafından keşfedilen en önemli 10 özneliğin, diğeri veri kümelerinde (sütun indeksiyle gösterilen) sınıflandırma doğruluğu analizi için sonuçlar raporlanmıştır. Sınıflandırma sonuçları için RF sınıflandırıcı kullanılmış ve 10 deneyin ortalaması hesaplanmıştır. Bu deneyde, kararlı öznelik kümelerinin, özellikle aynı hastalık grubundaki (İBH veya KK) veri kümelerinde sürekli olarak daha yüksek performans sağladığını göstermek amaçlanmıştır. Ancak, İBH ve KK'nin her ikisi de potansiyel olarak ortak faktörlere sahip bağırsak hastalıkları olduğundan ve İBH gelecekteki aşamalarda KK'ye dönüşme riski taşıdığından, tüm veri kümelerinde benzer sonuçlar keşfedilmiştir.



Şekil 4.7. Farklı veri kümeleri için Otokodlayıcı, U-Net, RPCA, mRMR ve RF yöntemleriyle elde edilen özneliliklerin veri kümeleri arası tutarlılığı.

4.1.5.3. Keşfedilen Öznitelik Kümelerinin Farklı Veri Kümeleri için Tutarlılık Analizi

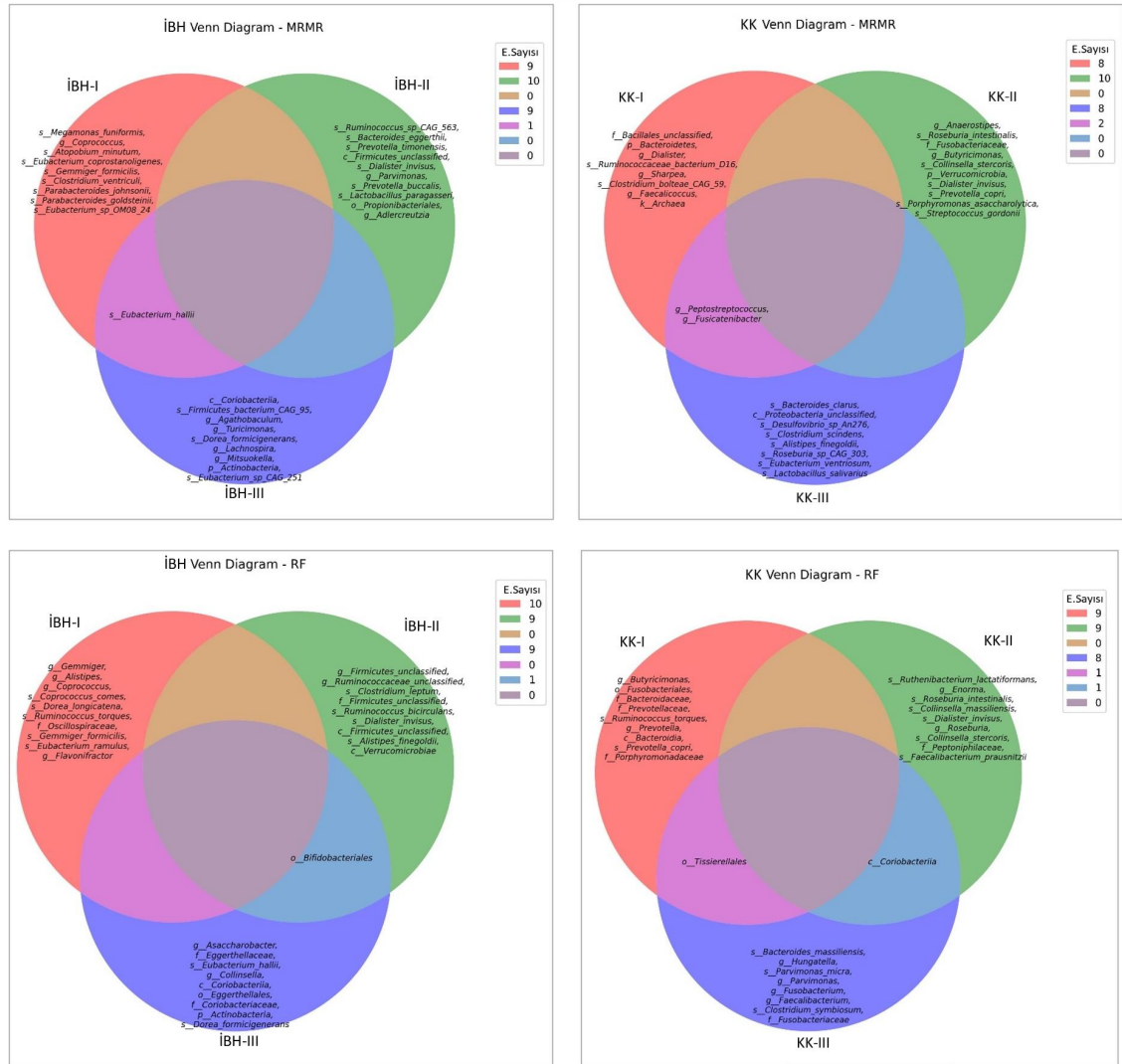
Bu analiz için, tutarlılık hesaplaması daha önceki bölümlerde açıklandığı gibi yapılmış, ancak hesaplama bu sefer veri kümesi çiftleri arasında gerçekleştirilmiştir. İki veri kümesi için keşfedilen onar öznitelik kümesi arasındaki benzerlik indeksi (KI) bulunmuş ve ardından genel tutarlılık (C_{avg}) sonuçları buna göre hesaplanmıştır. Deneylerin sonucunda (Şekil 4.7), önerilen yöntem ve RPCA tarafından bulunan özniteliklerin veri kümesi çiftleri boyunca uyumlu ve tutarlı olduğu, RF ve mRMR tarafından bulunan özniteliklerin ise hem aynı veri kümesi içinde hem de çapraz veri kümeleri arasında tutarsız olduğu görülmektedir.

4.1.5.4. İBH ve KK için Keşfedilen Potansiyel Biyobelirteçler

Bu bölümde, altı farklı veri kümesi için Otokodlayıcı, U-Net, RPCA, mRMR ve RF yöntemleriyle keşfedilen biyobelirteçler raporlanmıştır. Bu deneyde, her veri kümesi için modeller tüm veriler kullanılarak eğitilmiş ve on biyobelirteç içeren bir biyobelirteç kümesi elde edilmiştir. Her yöntem için, bu biyobelirteç kümelerinin aynı hastalıklar (İBH veya KK) için ortak olan eleman sayısı ve adları bildirilmiştir. Şekil 4.8’de Otokodlayıcı, U-Net ve RPCA ile keşfedilen biyobelirteçlerin İBH ve KK veri setlerinde oldukça yaygın olduğu gözlemlenirken, Şekil 4.9’da mRMR ve RF yöntemlerinin bu veri kümeleri için neredeyse hiçbir ortak biyobelirteç keşfedemediği fark edilmektedir. Ek olarak, üç ayrı İBH veri seti için Otokodlayıcı ile bulunan on biyobelirteçten beşinin, üç veri kümesi için de ortak olduğu, U-Net modeli ile bulunan ortak biyobelirteç sayısının ise sekiz olduğu gözlemlenmiştir. Benzer şekilde, KK veri setleri için Otokodlayıcı yöntemi ile keşfedilen biyobelirteçlerden altısı ve U-Net ile keşfedilen biyobelirteçlerden beşi, üç veri setinde de ortaktır. Önerilen yöntemle farklı veri kümelerinde keşfedilen ortak biyobelirteçlerin yüksek yaygınlık oranı (7/10, 8/10, vb.), önerilen yöntemin farklı veri kümeleri için benzer sonuçlar verebileceği, bu belirteçlerin potansiyel biyobelirteçler olabileceği ve ilgili hastalıkların (İBH ve KK) tanı ve tedavisinde kullanılabileceği inancını desteklemektedir.



Şekil 4.8. İBH ve KK veri kümelerinde Otokodlayıcı, U-Net ve RPCA yöntemleri ile elde edilen 10 biyobelirteç ve üç veri kümesi için ortak biyobelirteçler.



Şekil 4.9. İBH ve KK veri kümelerinde mRMR ve RF yöntemleri ile elde edilen 10 biyobelirteç ve üç veri kümesi için ortak biyobelirteçler.

Çalışmamızda Otokodlayıcı yöntemi kullanılarak üç ayrı İBH veri kümesi üzerinde gerçekleştirilen analizler sonucunda, yaygın olarak keşfedilen biyobelirteçlerin, İBH gelişim mekanizmalarıyla ilişkili olduğu bulunmuştur [167, 168]. Mevcut literatür, sonuçlarımızla tutarlı şekilde İBH'li bireylerde bağırsak mikrobiyotasının filum düzeyinde bakteriyel filum *Actinobacteria* [168–170] tarafından domine edildiğini ve daha az ölçüde *Bacteroidetes* [171,172] (*Bacteroidales* [173,174], *Bacteroides* [173,175]) içerdiğini bildirmektedir.

Aynı şekilde, U-Net yöntemi ile İBH veri setlerinde keşfedilen yaygın biyobelirteçler olan sınıf düzeyindeki *Bacteroidia* [176, 177] ve familya düzeyindeki *Bacteroidacea* [178–180] bolluğunun, İBH hastalarında sağlıklı kontrollere göre azaldığı bilinmektedir.

Ayrıca, *Streptococcus* [181, 182] ve *Methanomassiliicoccus*'un [183, 184] değişen bolluk düzeylerinin İBH ile ilişkili olduğu bilinmesine rağmen, keşfedilen biyobelirteçler arasında yer alan *Streptococcus viridans* ve *Candidatus methanomassiliicoccus intestinalis* ile ilgili literatürde doğrudan tür düzeyinde bir çalışma bulunamamıştır.

Çalışmamız tarafından tanımlanan KK biyobelirteçleri arasında *Firmicutes*, *Bacteroidetes*, *Clostridiales*, *Actinobacteria*, *Clostridia* ve *Lachnospiraceae* yer almaktadır. Önerilen çalışma ile keşfedilen biyobelirteç adaylarının, literatürde bulunan birçok çalışmada, KK ile yüksek oranda ilişkili olduğu raporlanmıştır [86, 185, 186]. Önceki çalışmalarla uyumlu olarak, sağlıklı bireylerde en belirgin şubeler arasında olan *Firmicutes* [187] (*Clostridia* [188, 189], *Clostridiales* [190]) ve *Bacteroidetes*'in [191], KK ile ilişkili biyobelirteçler olduğu söylenebilir [192]. Ek olarak, çeşitli çalışmalar, KK'li hastalarda *Actinobacteria* [193] ve *Lachnospiraceae* [194] bolluk düzeylerinde değişiklikler gözlemlendiğini kanıtlamıştır [195, 196].

5. BÖLÜM

TARTIŞMA, SONUÇ ve ÖNERİLER

5.1. Tartışma

Bağırsak mikrobiyomu bileşimi ile çeşitli hastalıklar arasındaki karmaşık ilişki, çok sayıda bilimsel araştırmanın odak noktası olmuştur, ancak erken ve invaziv olmayan teşhis için bu korelasyonu kullanmanın gerçek potansiyeli hala başlangıç aşamasındadır. Böyle bir yaklaşımın vaadi, hastalıkları erken bir aşamada tespit etme yeteneğinde yatmaktadır ve bu da potansiyel olarak, hasta geri bildirimlerinin daha iyi olmasına ve sağlık hizmeti maliyetlerinin azaltılmasına yol açacaktır. Bağırsak mikrobiyomundaki değişiklikleri yüksek duyarlılık ve özgüllükle tespit edebilme yeteneği, bu potansiyeli klinik uygulamalara adapte edebilmek için kritik bir faktördür. Mikrobiyom tabanlı bir tarama testi, yalnızca bu katı kriterlerle rutin teşhis prosedürlerine entegre edilebilir ve ancak bu koşullarda mevcut teşhis araçlarının etkinliğini desteklemekle kalmayıp, aynı zamanda iyileştirilmesini sağlayabilir. Thomas ve ark. tarafından yapılan çığır açan çalışma [197], bu potansiyelin bir kanıtıdır. Araştırmaları, dışkı örneklerinden elde edilen metagenomik verilerin, dışkıda gizli kan testi gibi geleneksel invaziv olmayan testlerin performansıyla rekabet edebileceğini, hatta bazı durumlarda onları geçebileceğini göstermiştir. Bu tür bulgular, invaziv olmayan tanılama yaklaşımı için bir paradigma değişiminin göstergesidir. Bu nedenle, tanılama performansındaki herhangi bir teknik iyileştirme, bağırsak mikrobiyomu analizinin potansiyelini ortaya çıkaracaktır. Sonuçlarımız, bağırsak mikrobiyomunu modelleme, hastalık sınıflandırma ve biyobelirteç keşfi görevleri için farklı modeller geliştirme girişimlerinin, hastalık imzalarını tutarlı bir şekilde keşfedebileceğini ve uygulanabilir tanı alternatifleri ortaya çıkarabileceğini göstermektedir. Ayrıca metagenomik verilerin tam potansiyelinin,

hastalık sınıflandırması, biyobelirteç keşfi ve tanılama işlevleri özelinde göz ardı edilmiş olabileceği ve daha gelişmiş matematiksel modellemeyle, potansiyeli ortaya çıkarmak için gerekli çabanın değerli olduğu söylenebilir. Makine öğrenimindeki ilerlemeler sayesinde, bağırsak mikrobiyomu verilerine dayalı hastalık tespitindeki gelişmeler, yazılım odaklı invaziv olmayan tanılama olarak klinik uygulamalara çevrilebilir. Moleküler tanılamadaki mevcut gelenek, çoğunlukla farklı hastalıklar için özel olarak in vitro test geliştirilmesini gerektirir. Buna karşılık, metagenomik veriler üzerine inşa edilen temel modeller, üzerinde çeşitli hastalıklar için ince ayar yapıldığında, tek bir testin çok çeşitli koşulları tanımlayabileceği birleşik yazılım platformlarının yaygınlaşmasını kolaylaştırabilir. Ayrıca, bu modeller invaziv olmayan evde tarama için kullanılabilir ve proaktif önleyici tıbbın temel bir bileşeni haline gelebilir.

Ayrıca çalışmamızda, derin öğrenmeye dayalı bir insan bağırsak mikrobiyomu modellemesi önerilmiş ve bu modellerle ilgili gürültülü bileşenlerin İBH ve KK veri kümeleri arasında ayırmacı olan gürbüz biyobelirteçlere karşılık geldiği gösterilmiştir. Bu denetimsiz strateji ile popüler denetimli yaklaşımlara kıyasla daha tutarlı keşifler sağlanırken, daha önce önerilen doğrusal muadillerine kıyasla daha üstün ve daha sağlam bir hastalık sınıflandırma performansına erişilmiştir. Bir hastalığın daha iyi anlaşılmasının, daha iyi tanı ve tedaviye yol açabileceği göz önünde bulundurularak, bağırsak mikrobiyomundaki hastalıkla ilgili olası biyobelirteçler araştırılmış ve bunların tanı ve tedavide kullanılabileceği önerilmiştir [87]. Mikrobiyal topluluklar söz konusu olduğunda, yapıları ve işlevleri çeşitli ortamlardaki ekolojik etkileşimlere ve mikrobiyal bağlantılara (karşılıklık, rekabet, sinerji vb.) büyük ölçüde bağımlıdır [198]. Bu nedenle, bu faktörlerin temel taksonların tanımlanmasıyla uğraşırken dikkate alınması gerektiği unutulmamalıdır. Ek olarak, mikrobiyomdan keşfedilen biyobelirteçlerin hastalıkların teşhis ve tedavisinde kullanılabilmesi için, bağırsak mikrobiyomundaki değişiklikleri yüksek hassaslık ve özgüllükle tespit edebilmeleri gerekir. Derin öğrenmenin metagenomik çalışmalara entegre edilmesi, araştırmacıların mikrobiyal toplulukların karmaşıklıklarında gezinmesine ve daha önce anlaşılması zor ilişkileri ortaya çıkarmasına olanak tanıyan yeni bir alan sağlar. Fakat, yenilikçi derin öğrenme çalışmalarında [141, 150] bile, aynı hastalıklar için özdeş veri kümeleri kullanılmasına rağmen, farklı yöntemlerle farklı sonuçlar (biyobelirteçler) elde edilebileceği açıktır.

Tutarlı olmayan biyobelirteç keşfi ve tekrarlanabilirliğin eksikliği, mikrobiyom biliminin klinik ve biyoteknolojik uygulamalara aktarılmasının önündeki en büyük engel olarak düşünülebilir. Bu amaçla, metagenomik veri setlerinde İBH ve KK ile ilişkili gürbüz biyobelirteç keşfine yönelik yeni bir yaklaşım benimsenmiştir. Otokodlayıcı modeli, metagenomik girdi verilerinin bir temsilini verimli bir şekilde öğrenerek, onu olabildiğince sadık bir şekilde yeniden yapılandırmayı ve temel ayırt edici biyolojik belirteçleri vurgulamayı amaçlamıştır. Böylece, geleneksel yaklaşımlar tarafından gözden kaçırılmış olabilecek biyobelirteçlerin keşfini kolaylaştırmış ve İBH ve KK ile ilişkili mikrobiyal kompozisyona ilişkin anlayışımızı geliştirmiştir. Veri kümelerindeki mekânsal ilişkileri yakalama becerisiyle bilinen U-Net mimarisi, metagenomik veri kümelerinin karmaşık yapısını analiz etme konusunda oldukça etkili olmuştur. Bu yaklaşımlar, yalnızca ayırt edici anahtar taksonları belirlemekle kalmayıp, aynı zamanda hastalık gelişiminde önemli rol oynayan mikrobiyal konfigürasyonlara da ışık tutmuştur.

Bulgularımız, çalışmamızın metagenomik veri kümelerinde biyobelirteç keşfinin gürbüzlüğüne katkıda bulunduğunu ve bağırsak mikrobiyomu ile İBH ve KK gibi hastalıklar arasındaki karmaşık ilişkileri çözmek için güçlü bir araç sunduğunu göstermektedir. Ek olarak, bu çalışmanın sonuçları, bağırsak mikrobiyomu veri kümelerinden tutarlı biyobelirteçler bulma konusunda, denetimsiz yöntemlerin denetimli yöntemlerden önemli ölçüde daha başarılı olduğunu göstermiştir. Bu sonuç, denetimsiz yöntemlerin, hastalık bağlamında aşırı uyuma çok yatkın bir süreç içermesi nedeniyle şaşırtıcı değildir. Çalışmamızda bu olguyu destekler şekilde mRMR ve RF yöntemleri, İBH ve KK veri kümeleri için herhangi bir ortak biyobelirteç keşfetmeyi başaramamıştır. Farklı şekilde düzenlenmedikçe, denetimli öznelik seçiminin metagenomik veri kümelerine pratik olarak uygulanamayacağı sonucuna varılabilir.

Temsili öğrenme, önceden eğitilmiş modeller veya temel modeller üzerinde, denetimsiz veya yarı denetimli öğrenme yaklaşımları aracılığıyla verilerin genel özelliklerini yakalamayı ifade eder. Bu paradigmaya göre, belirli görevlerin nispeten küçük alt kümeleri genel özellikleri sıfırdan çıkarmak için yetersiz kalabilir. Böylece tanımlayıcı özelliklerin aktarılması ve daha ayrımcı istatistiklerin öğrenilmesi daha iyi bir derin öğrenme stratejisi olarak benimsenebilir. Otokodlayıcı sinir ağları veya U-Net gibi darboğaz mimarileri aracılığıyla doğrusal olmayan boyut azaltma,

bu yaklaşımla uyumlu bir metodoloji olarak düşünülebilir. Daha düşük boyutlu gizli temsiller veya yeniden yapılandırılmış/gürültüden arındırılmış veriler kullanan derin öğrenme sınıflandırıcılarının daha sağlam ve doğru sınıflandırma sonuçları sağladığı birçok çalışma ile doğrulanmıştır. Bu çalışmadaki amacımız, bu fikri bağırsak mikrobiyomunun taksonomik manzarasına uyarlamaktır. Birleştirilmiş bir bağırsak metagenom verisi kohortu üzerinde denetimsiz ön eğitimin, geleneksel sınıflandırma algoritmalarının sınıflandırma performansını iyileştirebildiğini gösterdik. Bu çalışmanın temel katkılarından biri olan bu bulgu, bağırsak mikrobiyomu çalışmaları için, zorlukların çok yönlü olması sebebiyle oldukça önemlidir. Örneğin, bağırsak metagenomu verileri genellikle kohort başına onlarca, yüzlerce hatta çoğunlukla binlerce mertebesinde üretilmiştir. Bu rejimin getirdiği pratik ve teknik sınırlamalar olduğundan, az örnek-yüksek boyut sorunu, hipotezleri test etme zorluklarını da beraberinde getirir. Veriler yalnızca kıt değil, aynı zamanda yüksek boyutludur ve bu da onu "boyutluluğun laneti" olgusunun klasik bir örneği haline getirir. Bu sebeple, doğru koşullar sağlanmadığı sürece sağlık ile ilişkili biyobelirteçlerin gözden kaçabileceği ve doğru şekilde keşfedilemeyeceğini iddia etmek mümkündür. Bu nedenle, hem hastalık-mikrobiyom ilişkilerinin keşfi hem de bağırsak metagenomunun tanı performansının ortaya çıkarılması için önceden eğitilmiş modeller, nispeten küçük çalışma grupları için daha güçlü modelleme sunabilir. Temsili öğrenmenin gücünün fark yarattığı yer burasıdır. Denetimsiz ön eğitim ile etiketlenmemiş büyük miktardaki verinin gücünden yararlanılarak bir zorluğu avantaja dönüştürmek mümkündür. Bu çalışmanın diğer ana katkısı, temsili öğrenmeyi bağırsak mikrobiyomu verilerine uygulanabilir hale getirme becerisidir. Yapılan alt örnekleme deneyleri, son derece küçük alt gruplar için hastalık vakalarını sağlıklı kontrollerden bazı zayıf ancak istatistiksel olarak anlamlı sınıflandırmalarla ayırmanın mümkün olduğunu göstermiştir.

5.2. Sonuç ve Öneriler

Söz konusu çalışma ile bağırsak mikrobiyomunun taksonomik modellemesinin, belirli hastalıklar için sınıflandırma performansında iyileşmeye yol açıp açmayacağını araştırılmış, ayrıca birtakım hastalıklar ile ilişkili kararlı biyobelirteçlerin keşfi için yeni yaklaşımlar geliştirilmiştir. Önerilen yaklaşımlar, verilerdeki kalıpları ve ilişkileri

yakalama yeteneğinin iyileştirilmesi sayesinde hastalıklarla ilişkili mikrobiyal imzalara ilişkin anlayışı yenileme potansiyeline sahiptir. Mevcut yöntemler göz önüne alındığında, kapsamlı bir veri kümesi üzerinde eğitilen modellerimizin, mikrobiyal toplulukların içsel hiyerarşik yapısından yararlanarak bu biyolojik karmaşıklığı, derin öğrenme modellerinin sınıflandırma görevleri için etkili bir şekilde kullanabildiği ve bunun sonucunda üstün sınıflandırma performansı elde edebildiği gösterilmiştir. Ek olarak bulgularımız, derin öğrenme mimarilerinin metagenomik veri kümelerindeki karmaşık kalıpları yakalamadaki etkinliğini ve böylece gelişmiş doğrulukla potansiyel gurbüz biyobelirteçleri tespit edebildiğini göstermektedir. Belirlenen biyobelirteçler yalnızca potansiyel invaziv olmayan tanı göstergeleri olarak hizmet etmekle kalmaz, aynı zamanda yaklaşımımızın klinik önemini vurgular şekilde hedefli terapötik müdahaleler için de yol açabilir.

Geliştirilen mimarilerin sinir ağı tabanlı çerçevesi, geleneksel yaklaşımlara göre açık bir avantaj sunmaktadır. Bu mimari sayesinde, alternatif farklı veri setleriyle çalışmayı kolaylaştırırken, aynı zamanda diğer modellerin baş/boyun modülü olarak diğer sinir ağı tabanlı yöntemlere de entegre edilebilir. Bu taşınabilir ve çok yönlülük özellikleri sayesinde, bir sağlık indeksi olarak hizmet etmek üzere uyarlanabilir ve genişletilebilir. Bulgularımız, transfer öğreniminin etkili kullanımının, modellerin ham ve işlenmemiş mikrobiyom verilerinde yakalanması mümkün olmayan ince ve karmaşık desenleri tanıma ve oldukça az örneğe sahip veri kümeleriyle sınıflandırma yapabilme yeteneğini geliştirdiğini kanıtlamıştır. Bu nedenle, istatistiksel olarak yetersiz veri kümeleri için bile birkaç atışlı bir hipotez test şeması önermek mümkün olabilir. Gelecekteki modeller, yeterli miktarda metagenomik veri ile hastalık sınıflandırması ve teşhisinde benzeri görülmemiş bir başarı sağlayabilir. Olası modellerin sıfır atış ve birkaç atış yetenekleri, daha önce keşfedilmemiş bağırsak mikrobiyomu-hastalık ilişkilerini de ortaya çıkarabilir. Bulgularımızın daha geniş etkileri, ilişkili bozuklukların geniş bir yelpazesinde bağırsak mikrobiyomuna dayalı invaziv olmayan hastalık teşhisini teşvik etmenin yanı sıra konak mikrobiyom etkileşimlerinin hassas modellemesini içerir. Mevcut çabamız, bağırsak metagenom verilerinin hastalıklarla ilişkisini belirlemek ve devam eden araştırmaları pratik tanı ve tedavi uygulamalarına dönüştürmek amacıyla işlenmesini kolaylaştırmaya ve bu sonuçların tekrarlanabilirliğini desteklemeye odaklanmaktadır.

Sonuçlarımız, sağlıklı kontroller ve farklı İBH ve KK kohortları kullanılarak oluşturulmuş bir dizi bağırsak metagenom veri kümesi üzerinde, derin modeller ile erişilen yeniden yapılandırma kaybının analizine vurgu yapmanın biyobelirteç keşfi için sağlam bir çerçeve sağladığını göstermektedir. Bu, keşif algoritmalarının tutarlılığını test etmek için yeterli bir ölçüt sağlarken, metodolojinin daha gerçekçi bir potansiyeli yalnızca gerçek bir doğrulama kohortu üzerinde ölçülebilir. Dahası, nispeten büyük kohortlarla (örneğin, mevcut veri kümelerinden birkaç kat daha büyük) çalışan eğitim modelleri, metagenomik modellemenin temel yeteneklerinin ve bilinmeyen farklı kapasitelerinin (varsa) ortaya çıkmasını mümkün kılacaktır. Araştırmamız yalnızca kısıtlı hastalıklara odaklanmaktadır. Genelleştirilebilirliği sağlamak için bu modellerin daha geniş ve çeşitli bir hastalık verisi yelpazesi kullanılarak eğitilmesi esastır. Bu çalışma için gelecekteki yönler, bağırsak metagenomlarının kapsamlı ve daha büyük kohortlarında eğitilmiş modeller kullanarak ince ayar, az atışlı öğrenme ve sıfır atışlı yeteneklere daha derinlemesine eğilmeyi içerir. Bu çalışmada özetlenen ilkelerin, derin öğrenme modelleri kullanılarak geniş süper kohortlarda uygulanmasının umut verici bir yol olması beklenmektedir. Böylece bu yaklaşımlar, mikrobiyom alanındaki denetimsiz öğrenme yeteneklerinin daha kapsamlı bir şekilde değerlendirilmesine olanak sağlayacaktır.

Çalışmamız metagenomik biyobelirteç keşfinde önemli bir ilerlemeyi temsil etse de gelecekteki araştırmalar için farklı yollar vardır. Öncelikle, metatranskriptomik ve metabolomik gibi çoklu omik verilerin dahil edilmesi, hastalık koşullarında bağırsak mikrobiyomunun işlevsel yönleri hakkında daha kapsamlı bir anlayış sağlayabilir. Transfer öğrenme ve topluluk yöntemlerinin dahili, farklı hasta popülasyonları ve farklı veri kümeleri için model genellemesini daha da geliştirebilir. Ayrıca, derin öğrenme modellerinin yorumlanabilirliği en büyük zorluk olmaya devam etmektedir. Gelecekteki çabalarda, yeniden yapılandırılmış verilerde belirlenen farklılıkların biyolojik önemini yorumlamak için yöntemlerin geliştirilmesi, hesaplamalı bulguları klinik olarak eyleme geçirilebilir içgörülere dönüştürmek için kritik olacaktır. Alan bu temeller üzerine inşa edilerek, yeniden yapılandırılmış metagenomik veriler perspektifinden hastalıkları yönetmek için daha hedefli ve kişiselleştirilmiş yaklaşımlara doğru ilerleyebilir. Ek olarak, DL modelleri ile yeniden yapılandırılmış veriler üzerinde öznetelik seçimi ve boyut azaltma için gelişmiş tekniklerin araştırılması, biyobelirteç keşfinin verimliliğini

artırabilir. Sonuç olarak, gelişmiş derin öğrenme yaklaşımlarının entegrasyonu ve yorumlanabilirliğe odaklanma, şüphesiz metagenomik alanında biyobelirteç keşfi ve hastalık tespiti uygulamalarını ileriye taşıyacak ve hastalıklar için yeni içgörüler ve potansiyel terapötik hedefler sunacaktır.



KAYNAKÇA

1. Alshawaqfeh, M., Bashaireh, A., Serpedin, E., and Suchodolski, J., 2017. Consistent metagenomic biomarker detection via robust PCA, **Biology Direct**, **12**(1).
2. Jiménez, B.G., Muñoz, J., Cabello, S., Medina, J., and Wilkinson, M.D., 2021. Predicting microbiomes through a deep latent space, **Bioinformatics**, **37**(10):1444–1451.
3. Namkung, J., 2020. Machine learning methods for microbiome studies, **J Microbiol**, **58**(3):206–216.
4. Le, V., Quinn, T.P., Tran, T., and Venkatesh, S., 2020. Deep in the Bowel: Highly Interpretable Neural Encoder-Decoder Networks Predict Gut Metabolites from Gut Microbiome, **BMC Genomics**, **21**(S4).
5. Boucheham, A. and Batouche, M., 2014. Robust biomarker discovery for cancer diagnosis based on meta-ensemble feature selection, *Science and Information Conference*, 452–560, IEEE.
6. Azencott, C.A., 2020. Machine learning tools for biomarker discovery, *Sorbonne University UPMC*, Computer science bibliography.
7. Lim, J., Bang, S., and Kim, J.v., 2019. Integrative Deep Learning for Identifying Differentially Expressed (DE) Biomarkers, **Computational and Mathematical Methods in Medicine**, **2019**:1–10.
8. Van Malderen, K. and De Winter, B.Y.v., 2020. Volatomics in inflammatory bowel disease and irritable bowel syndrome, **EBioMedicine**, **54**:102725.
9. Kang, D.W., Adams, J.B., and Coleman, D.M.v., 2019. Long-term benefit of Microbiota Transfer Therapy on autism symptoms and gut microbiota, **Scientific Reports**, **9**(1).
10. Papoutsoglou, G., Tarazona, S., and Lopes, M.B.v., 2023. Machine learning approaches in microbiome research: challenges and best practices, **Frontiers in Microbiology**, **14**.

11. Abeel, T. and Helleputte, T., 2010. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, **Bioinform.**, **26**(3):392–398.
12. Alshawaqfeh, M., Bashaireh, A., Serpedin, E., and Suchodolski, J., 2017. Reliable Biomarker discovery from Metagenomic data via RegLRSD algorithm, **BMC Bioinform.**, **18**(1):328.
13. Cheung, H. and Yu, J., 2021. Machine learning on microbiome research in gastrointestinal cancer, **Journal of Gastroenterology and Hepatology**, **36**(4):817–822.
14. Manichanh, C., Borruel, N., Casellas, F., and Guarner, F., 2012. The gut microbiota in IBD, **Nature Reviews Gastroenterology & Hepatology**, **9**(10):599–608.
15. Roy, S. and Trinchieri, G., 2017. Microbiota: a key orchestrator of cancer therapy, **Nature Reviews Cancer**, **17**(5):271–285.
16. Turnbaugh, P.J., Ley, R.E., and Mahowald, M.A.v., 2006. An obesity-associated gut microbiome with increased capacity for energy harvest, **Nature**, **444**(7122):1027–1031.
17. Qin, J., Li, Y., and Cai, Z.e.a., 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes, **Nature**, **490**(7418):55–60.
18. Chuang, H., Lee, E., Liu, Y., Lee, D., and Ideker, T., 2007. Network-based classification of breast cancer metastasis, **Molecular Systems Biology**, **3**(1).
19. van 't Veer, L.J., Dai, H., and van de Vijver, M.J.v., 2002. Gene expression profiling predicts clinical outcome of breast cancer, **Nature**, **415**(6871):530–536.
20. Wang, Y., Klijn, J.G., and Zhang, Y.v., 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, **The Lancet**, **365**(9460):671–679.
21. Turnbaugh, P.J. and Ley, R.E.v., 2007. The Human Microbiome Project, **Nature**, **449**(7164):804–810.

22. Shreiner, A.B., Kao, J.Y., and Young, V.B., 2015. The gut microbiome in health and in disease, **Current Opinion in Gastroenterology**, **31**(1):69–75.
23. Lozupone, C.A., Stombaugh, J.I., and vd., 2012. Diversity, stability and resilience of the human gut microbiota, **Nature**, **489**(7415):220–230.
24. Sommer, F. and Bäckhed, F., 2013. The gut microbiota — masters of host development and physiology, **Nature Reviews Microbiology**, **11**(4):227–238.
25. Nicholson, J.K., Holmes, E., and Kinross, J.v., 2012. Host-Gut Microbiota Metabolic Interactions, **Science**, **336**(6086):1262–1267.
26. Segata, N., Izard, J., and Waldron, L.v., 2011. Metagenomic biomarker discovery and explanation, **Genome Biology**, **12**(6):R60.
27. Yu, Y.N. and Fang, J.Y., 2015. Gut Microbiota and Colorectal Cancer, **Gastrointestinal Tumors**, **2**(1):26–32.
28. Quaglio, A.E.V. and Grillo, T.G.v., 2022. Gut microbiota, inflammatory bowel disease and colorectal cancer, **World Journal of Gastroenterology**, **28**(30):4053–4060.
29. Scher, J.U. and Abramson, S.B., 2011. The microbiome and rheumatoid arthritis, **Nature Reviews Rheumatology**, **7**(10):569–578.
30. Ferreira, A.L., Choi, J., Ryou, J., and Newcomer, E.P.v., 2023. Gut microbiome composition may be an indicator of preclinical Alzheimer's disease, **Science Translational Medicine**, **15**(700).
31. Jangi, S., Gandhi, R., Cox, L.M., and Li, N.v., 2016. Alterations of the human gut microbiome in multiple sclerosis, **Nature Communications**, **7**(1).
32. Sender, R., Fuchs, S., and Milo, R., 2016. Revised Estimates for the Number of Human and Bacteria Cells in the Body, **PLOS Biology**, **14**(8):e1002533.
33. Qin, J., Li, R., and Raes, J.v., 2010. A human gut microbial gene catalogue established by metagenomic sequencing, **Nature**, **464**(7285):59–65.

34. Abbas, M., Matta, J., and Le, T.v., 2019. Biomarker discovery in inflammatory bowel diseases using network-based feature selection, **PLOS ONE**, **14**(11):e0225382.
35. Yamamoto, M. and Matsumoto, S., 2016. Gut microbiota and colorectal cancer, **Genes and Environment**, **38**(1).
36. Xavier, R.J. and Podolsky, D.K., 2007. Unravelling the pathogenesis of inflammatory bowel disease, **Nature**, **448**(7152):427–434.
37. Mukhopadhyay, I., Hansen, R., and Meharg, C.v., 2015. The fungal microbiota of de-novo paediatric inflammatory bowel disease, **Microbes and Infection**, **17**(4):304–310.
38. Kostic, A.D., Xavier, R.J., and Gevers, D., 2014. The Microbiome in Inflammatory Bowel Disease: Current Status and the Future Ahead, **Gastroenterology**, **146**(6):1489–1499.
39. Vogtmann, E., Hua, X., and Zeller, G.v., 2016. Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing, **PLOS ONE**, **11**(5):e0155362.
40. Niloofa, R., De Zoysa, M., and Seneviratne, L., 2021. Autoantibodies in the diagnosis, prognosis, and prediction of colorectal cancer, **Journal of Cancer Research and Therapeutics**, **17**(4):819.
41. Zhu, J., Luo, J., and Ma, Y., 2023. Screening of serum exosome markers for colorectal cancer based on Boruta and multi-cluster feature selection algorithms, **Molecular & Cellular Toxicology**, **20**(2):343–351.
42. Launoy, G., Smith, T.C., Duffy, S.W., and Bouvier, V., 1997. Colorectal cancer mass-screening: Estimation of faecal occult blood test sensitivity, taking into account cancer mean sojourn time, **International Journal of Cancer**, **73**(2):220–224.

43. Nadeem, M.S., Kumar, V., Al-Abbasi, F.A., Kamal, M.A., and Anwar, F., 2020. Risk of colorectal cancer in inflammatory bowel diseases, **Seminars in Cancer Biology**, **64**:51–60.
44. Manandhar, I., Alimadadi, A., and Aryal, S.v., 2021. Gut microbiome-based supervised machine learning for clinical diagnosis of inflammatory bowel diseases, **American Journal of Physiology-Gastrointestinal and Liver Physiology**, **320**(3):G328–G337.
45. Waljee, A.K., Lipson, R., and vd., 2017. Predicting Hospitalization and Outpatient Corticosteroid Use in Inflammatory Bowel Disease Patients Using Machine Learning, **Inflammatory Bowel Diseases**, **24**(1):45–53.
46. Hassan, C., 2008. Computed tomographic colonography to screen for colorectal cancer, extracolonic cancer, and aortic aneurysm: model simulation with cost-effectiveness analysis, **Archives of Internal Medicine**, **168**(7):696.
47. Mulenga, M. and Abdul Kareem, S.v., 2021. Feature Extension of Gut Microbiome Data for Deep Neural Network-Based Colorectal Cancer Classification, **IEEE Access**, **9**:23565–23578.
48. García-Bilbao, A., Armañanzas, R., and Ispizua, Z.v., 2012. Identification of a biomarker panel for colorectal cancer diagnosis, **BMC Cancer**, **12**(1).
49. Tarallo, S., Ferrero, G., and Gallo, G.v., 2019. Altered Fecal Small RNA Profiles in Colorectal Cancer Reflect Gut Microbiome Composition in Stool Samples, **mSystems**, **4**(5).
50. Milosevic, I., Vujovic, A., and Barac, A.v., 2019. Gut-Liver Axis, Gut Microbiota, and Its Modulation in the Management of Liver Diseases: A Review of the Literature, **International Journal of Molecular Sciences**, **20**(2):395.
51. Slattery, J., Macfabe, D.F., and Frye, R.E., 2016. The Significance of the Enteric Microbiome on the Development of Childhood Disease: A Review of Prebiotic and Probiotic Therapies in Disorders of Childhood, **Clinical Medicine Insights: Pediatrics**, **10**:CMPed.S38338.

52. Wang, X.q. and Zhang, A.h.v., 2018. Gut microbiota as important modulator of metabolism in health and disease, **RSC Advances**, **8**(74):42380–42389.
53. Paul, B., Barnes, S., and Demark-Wahnefried, W.v., 2015. Influences of diet and the gut microbiome on epigenetic modulation in cancer and other diseases, **Clinical Epigenetics**, **7**(1).
54. Mars, R.A., Yang, Y., Ward, T., and Houtti, M.v., 2020. Longitudinal Multi-omics Reveals Subset-Specific Mechanisms Underlying Irritable Bowel Syndrome, **Cell**, **182**(6):1460–1473.e17.
55. Valitutti, F., Cucchiara, S., and Fasano, A., 2019. Celiac Disease and the Microbiome, **Nutrients**, **11**(10):2403.
56. Stephens, Z.D., Lee, S.Y., and Faghri, F.v., 2015. Big Data: Astronomical or Genomical?, **PLOS Biology**, **13**(7):e1002195.
57. Krause, T. and Wassan, J.T.v., 2021. Analyzing Large Microbiome Datasets Using Machine Learning and Big Data, **BioMedInformatics**, **1**(3):138–165.
58. Roy, G., Prifti, E., Belda, E., and Zucker, J.D., 2024. Deep learning methods in metagenomics: a review, **Microbial Genomics**, **10**(4).
59. Jagga, Z. and Gupta, D., 2015. Machine Learning for Biomarker Identification in Cancer Research – Developments Toward its Clinical Application, **Personalized Medicine**, **12**(4):371–387.
60. Hernández Medina, R. and Kutuzova, S.v., 2022. Machine learning and deep learning applications in microbiome research, **ISME Communications**, **2**(1).
61. Cortes, C. and Vapnik, V., 1995. Support-vector networks, **Machine Learning**, **20**(3):273–297.
62. Berkson, J., 1944. Application to the Logistic Function to Bio-Assay, **Journal of the American Statistical Association**, **39**(227):357.
63. Teixeira, M., Silva, F., and Ferreira, R.M.v., 2024. A review of machine learning methods for cancer characterization from microbiome data, **npj Precision Oncology**, **8**(1).

64. Topçuoğlu, B.D., Lesniak, N.A., Ruffin, M.T., Wiens, J., and Schloss, P.D., 2020. A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems, **mBio**, **11**(3).
65. Arabameri, A., Asemani, D., and Teymourpour, P., 2020. Detection of Colorectal Carcinoma Based on Microbiota Analysis Using Generalized Regression Neural Networks and Nonlinear Feature Selection, **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, **17**(2):547–557.
66. Pasolli, E., Truong, D.T., Malik, F., Waldron, L., and Segata, N., 2016. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights, **PLOS Computational Biology**, **12**(7):e1004977.
67. Reiman, D., Metwally, A.A., Sun, J., and Dai, Y., 2020. PopPhy-CNN: A Phylogenetic Tree Embedded Architecture for Convolutional Neural Networks to Predict Host Phenotype From Metagenomic Data, **IEEE Journal of Biomedical and Health Informatics**, **24**(10):2993–3001.
68. Li, P., Luo, H., Ji, B., and Nielsen, J., 2022. Machine learning for data integration in human gut microbiome, **Microbial Cell Factories**, **21**(1).
69. Turnbaugh, P.J., Hamady, M., Yatsunenko, T., and Cantarel, B.L.v., 2008. A core gut microbiome in obese and lean twins, **Nature**, **457**(7228):480–484.
70. Saeys, Y., Abeel, T., and Van de Peer, Y., 2008. Robust Feature Selection Using Ensemble Feature Selection Techniques, 313–325, Springer Berlin Heidelberg.
71. Venkatesh, B. and Anuradha, J., 2019. A Review of Feature Selection and Its Methods, **Cybernetics and Information Technologies**, **19**(1):3–26.
72. Hwang, T., Sicotte, H., Tian, Z., and Wu, B.v., 2008. Robust and efficient identification of biomarkers by classifying features on graphs, **Bioinformatics**, **24**(18):2023–2029.
73. Eraslan, G., Avsec, Z., Gagneur, J., and Theis, F.J., 2019. Deep learning: new computational modelling techniques for genomics, **Nature Reviews Genetics**, **20**(7):389–403.

74. Camacho, D.M. and Collins, K.M.v., 2018. Next-Generation Machine Learning for Biological Networks, **Cell**, **173**(7):1581–1592.
75. Zhu, Q., Huo, B., Sun, H., Li, B., and Jiang, X., 2020. Application of Deep Learning in Microbiome, **Journal of Artificial Intelligence for Medical Sciences**, **1**(1–2):23–29.
76. Lo, C. and Marculescu, R., 2019. MetaNN: accurate classification of host phenotypes from metagenomic data using neural networks, **BMC Bioinformatics**, **20**(S12).
77. Ling, W., Qi, Y., Hua, X., and Wu, M.C., 2021. Deep ensemble learning over the microbial phylogenetic tree (DeepEn-Phy), *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 470–477, IEEE.
78. Nguyen, T.H. and Prifti, E.v., 2018, Disease Classification in Metagenomics with 2D Embeddings and Deep Learning.
79. Metwally, A.A., Yu, P.S., and Reiman, D.v., 2019. Utilizing longitudinal microbiome taxonomic profiles to predict food allergy via Long Short-Term Memory networks, **PLOS Computational Biology**, **15**(2):e1006693.
80. Sharma, D. and Xu, W., 2021. phyLoSTM: a novel deep learning model on disease prediction from longitudinal microbiome data, **Bioinformatics**, **37**(21):3707–3714.
81. Sharma, D., Paterson, A.D., and Xu, W., 2020. TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction, **Bioinformatics**, **36**(17):4544–4550.
82. Eck, S.H., 2018. Challenges in data storage and data management in a clinical diagnostic setting, **LaboratoriumsMedizin**, **42**(6):219–224.
83. Wu, G., Zhao, N., Zhang, C., Lam, Y.Y., and Zhao, L., 2021. Guild-based analysis for understanding gut microbiome in human health and diseases, **Genome Medicine**, **13**(1).

84. Gupta, V.K., Kim, M., Bakshi, U., and Cunningham, K.Y.v., 2020. A predictive index for health status using species-level gut microbiome profiling, **Nature Communications**, **11**(1).
85. Oh, M. and Zhang, L., 2020. DeepMicro: deep representation learning for disease prediction based on microbiome data, **Scientific Reports**, **10**(1).
86. Chen, X., Zhu, Z., and Zhang, W.v., 2022. Human disease prediction from microbiome data by multiple feature fusion and deep learning, **iScience**, **25**(4):104081.
87. Forouzandeh, A., Rutar, A., Kalmady, S.V., and Greiner, R., 2022. Analyzing biomarker discovery: Estimating the reproducibility of biomarker sets, **PLOS ONE**, **17**(7):e0252697.
88. Candès, E.J., Li, X., Ma, Y., and Wright, J., 2011. Robust principal component analysis?, **Journal of the ACM**, **58**(3):1–37.
89. Lin, Z., Chen, M., and Ma, Y., 2010. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices, **CoRR**, **abs/1009.5055**.
90. Choi, R., Coyner, A., Kalpathy-Cramer, J., Chiang, M., and Campbell, J., 2020. Introduction to Machine Learning, Neural Networks, and Deep Learning, **Translational vision science & technology**, **17**(9(2)):14.
91. Goodfellow, I.J., Bengio, Y., and Courville, A.C., 2016. Deep Learning, Adaptive computation and machine learning, MIT Press.
92. Ser, G. and Bati, C.T., 2019. Derin Sinir Ağları ile En İyi Modelin Belirlenmesi: Mantar Verileri Üzerine Keras Uygulaması, **Yüzüncü Yıl Üniversitesi Tarım Bilimleri Dergisi**, **29**(3):406–417.
93. Ramachandran, P., Zoph, B., and Le, Q.V., 2017, Searching for Activation Functions.
94. Vakalopoulou, M., Christodoulidis, S., Burgos, N., Colliot, O., and Lepetit, V., 2023. Deep Learning: Basics and Convolutional Neural Networks (CNNs), 77–115, Springer US.

95. Malhathkar, S. and S, T., 2022. Deep Learning for Time Series Forecasting – With a focus on Loss Functions and Error Measures, *2022 IEEE World Conference on Applied Intelligence and Computing (AIC)*, volume 13, 646–651, IEEE.
96. Rumelhart, D.E. and McClelland, J.L., 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations, The MIT Press.
97. Kingma, D.P. and Ba, J., 2014, Adam: A Method for Stochastic Optimization.
98. Duchi, J., Hazan, E., and Singer, Y., 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, **Journal of Machine Learning Research**, **12**:2121–2159.
99. Durairajah, V., Gobe, S., and Muneer, A., 2018. Automatic Vision Based Classification System Using DNN and SVM Classifiers, *2018 3rd International Conference on Control, Robotics and Cybernetics (CRC)*, volume 5, 6–14, IEEE.
100. Bank, D., Koenigstein, N., and Giryas, R., 2021, Autoencoders.
101. Li, P., Pei, Y., and Li, J., 2023. A comprehensive survey on design and application of autoencoder in deep learning, **Applied Soft Computing**, **138**:110176.
102. Karpinski, M., Khoma, V., Dudvkevych, V., Khoma, Y., and Sabodashko, D., 2018. Autoencoder Neural Networks for Outlier Correction in ECG- Based Biometric Identification, *2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS)*, IEEE.
103. Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.A., 2008. Extracting and composing robust features with denoising autoencoders, *Proceedings of the 25th international conference on Machine learning - ICML '08*, ICML '08, 1096–1103, ACM Press.
104. Coates, A., Ng, A.Y., and Lee, H., 2011. An Analysis of Single-Layer Networks in Unsupervised Feature Learning, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort*

Lauderdale, USA, April 11-13, 2011, volume 15 of JMLR Proceedings, 215–223, JMLR.org.

105. Kingma, D.P. and Welling, M., 2013, Auto-Encoding Variational Bayes.
106. Binbusayyis, A. and Vaiyapuri, T., 2021. Unsupervised deep learning approach for network intrusion detection combining convolutional autoencoder and one-class SVM, **Applied Intelligence**, **51**(10):7094–7108.
107. Ronneberger, O., Fischer, P., and Brox, T., 2015, U-Net: Convolutional Networks for Biomedical Image Segmentation.
108. Liu, H., Hu, D., Li, H., and Oguz, I., 2023. Medical Image Segmentation Using Deep Learning, 391–434, Springer US.
109. Zhang, Y., Qiu, Z., and vd., 2018. Fully Convolutional Adaptation Networks for Semantic Segmentation, **CoRR**, **abs/1804.08286**.
110. He, K., Zhang, X., Ren, S., and Sun, J., 2016. Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778, IEEE Computer Society.
111. Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K.Q., 2016, Densely Connected Convolutional Networks.
112. Oktay, O., Schlemper, J., Folgoc, L.L., and Lee, M.v., 2018. Attention U-Net: Learning Where to Look for the Pancreas, **arXiv**.
113. Wenzel, M., 2023. Generative Adversarial Networks and Other Generative Models, 139–192, Springer US.
114. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., and Xu, B.v., 2014, Generative Adversarial Networks.
115. ALTUN, S. and TALU, M.F., 2019. Aynı Şartlar Altında Farklı Üretici Çekişmeli Ağların Karşılaştırılması, *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 1–6, IEEE.

116. Radford, A., Metz, L., and Chintala, S., 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, *Y. Bengio and Y. LeCun, Eds.: 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.
117. Liang, X., Hu, Z., Zhang, H., Gan, C., and Xing, E.P., 2017, Recurrent Topic-Transition GAN for Visual Paragraph Generation.
118. Donahue, C., McAuley, J.J., and Puckette, M.S., 2019. Adversarial Audio Synthesis, *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net*.
119. Sidhom, J.W., Larman, H.B., Pardoll, D.M., and Baras, A.S., 2021. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires, **Nature Communications**, **12**(1).
120. Mahapatra, D., Bozorgtabar, B., and Garnavi, R., 2019. Image super-resolution using progressive generative adversarial networks for medical image analysis, **Computerized Medical Imaging and Graphics**, **71**:30–39.
121. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., and Jackel, L.D., 1989. Backpropagation Applied to Handwritten Zip Code Recognition, **Neural Computation**, **1**(4):541–551.
122. LeCun, Y., Bengio, Y., and Hinton, G., 2015. Deep learning, **Nature**, **521**(7553):436–444.
123. Thrun, S., 1995. Is Learning The n-th Thing Any Easier Than Learning The First?, *D.S. Touretzky, M. Mozer, and M.E. Hasselmo, Eds.: Advances in Neural Information Processing Systems 8, NIPS, Denver, CO, USA, November 27-30, 1995*, 640–646, MIT Press.
124. Pan, S.J. and Yang, Q., 2010. A Survey on Transfer Learning, **IEEE Transactions on Knowledge and Data Engineering**, **22**(10):1345–1359.

125. Hoarfrost, A., Aptekmann, A., Farfañuk, G., and Bromberg, Y., 2022. Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter, **Nature Communications**, **13**(1).
126. David, M.M., Tataru, C., Pope, Q., and Baker, L.J.v., 2022. Revealing General Patterns of Microbiomes That Transcend Systems: Potential and Challenges of Deep Transfer Learning, **mSystems**, **7**(1).
127. Tataru, C.A. and David, M.M., 2020. Decoding the language of microbiomes using word-embedding techniques, and applications in inflammatory bowel disease, **PLOS Computational Biology**, **16**(5):e1007859.
128. Turney, P., 1995. Technical note: Bias and the quantification of stability, **Machine Learning**, **20**(1–2):23–33.
129. Sen, R., Mandal, A.K., and Chakraborty, B., 2021. A Critical Study on Stability Measures of Feature Selection with a Novel Extension of Lustgarten Index, **Machine Learning and Knowledge Extraction**, **3**(4):771–787.
130. Kalousis, A., Prados, J., and Hilario, M., 2006. Stability of feature selection algorithms: a study on high-dimensional spaces, **Knowledge and Information Systems**, **12**(1):95–116.
131. Kuncheva, L.I., 2007. A stability index for feature selection, *V. Devedzic, Ed.; IASTED International Conference on Artificial Intelligence and Applications, part of the 25th Multi-Conference on Applied Informatics, Innsbruck, Austria, February 12-14, 2007*, 421–427, IASTED/ACTA Press.
132. Asgari, E., Garakani, K., McHardy, A.C., and Mofrad, M.R.K., 2018. MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples, **Bioinformatics**, **34**(13):i32–i42.
133. Wang, Y., Bhattacharya, T., and Jiang, Y.v., 2020. A novel deep learning method for predictive modeling of microbiome data, **Briefings in Bioinformatics**, **22**(3).

134. Vangay, P., Hillmann, B.M., and Knights, D., 2019. Microbiome Learning Repo (ML Repo): A public repository of microbiome regression and classification tasks, **GigaScience**, **8**(5).
135. Lee, S.J. and Rho, M., 2022. Multimodal deep learning applied to classify healthy and disease states of human microbiome, **Scientific Reports**, **12**(1).
136. Oudah, M. and Henschel, A., 2018. Taxonomy-aware feature engineering for microbiome classification, **BMC Bioinformatics**, **19**(1).
137. Zhang, S., Tong, H., Xu, J., and Maciejewski, R., 2019. Graph convolutional networks: a comprehensive review, **Computational Social Networks**, **6**(1).
138. Li, B., Zhong, D., Jiang, X., and He, T., 2021. TopoPhy-CNN: Integrating Topological Information of Phylogenetic Tree for Host Phenotype Prediction From Metagenomic Data, *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, volume 101, 456–461, IEEE.
139. Oh, M. and Zhang, L., 2023. DeepGeni: deep generalized interpretable autoencoder elucidates gut microbiota for better cancer immunotherapy, **Scientific Reports**, **13**(1).
140. García-Jiménez, B., Muñoz, J., Cabello, S., Medina, J., and Wilkinson, M.D., 2020. Predicting microbiomes through a deep latent space, **Bioinformatics**, **37**(10):1444–1451.
141. Shen, Y., Zhu, J., Deng, Z., Lu, W., and Wang, H., 2023. EnsDeepDP: An Ensemble Deep Learning Approach for Disease Prediction Through Metagenomics, **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, **20**(2):986–998.
142. Grazioli, F. and Siarheyev, R.v., 2022. Microbiome-based disease prediction with multimodal variational information bottlenecks, **PLOS Computational Biology**, **18**(4):e1010050.

143. Li, Y., Xie, G., Zha, Y., and Ning, K., 2023. GAN-GMHI: a generative adversarial network with high discriminative power for microbiome-based disease prediction, **Journal of Genetics and Genomics**, **50**(12):1026–1028.
144. Choi, J.M., Ji, M., Watson, L.T., and Zhang, L., 2023. DeepMicroGen: a generative adversarial network-based method for longitudinal microbiome data imputation, **Bioinformatics**, **39**(5).
145. Rong, R. and Jiang, S.v., 2021. MB-GAN: Microbiome Simulation via Generative Adversarial Network, **GigaScience**, **10**(2).
146. Reiman, D. and Dai, Y., 2020. Using Conditional Generative Adversarial Networks to Boost the Performance of Machine Learning in Microbiome Datasets, *Proceedings of the 1st International Conference on Deep Learning Theory and Applications*, 103–110, SCITEPRESS - Science and Technology Publications.
147. Hu, W., Yang, X., Wang, L., and Zhu, X., 2023. MADGAN: A microbe-disease association prediction model based on generative adversarial networks, **Frontiers in Microbiology**, **14**.
148. Jiang, X. and Zhao, J.v., 2020. A Generative Adversarial Network Model for Disease Gene Prediction With RNA-seq Data, **IEEE Access**, **8**:37352–37360.
149. Ditzler, G., Morrison, J.C., Lan, Y., and Rosen, G.L., 2015. Fizzy: feature subset selection for metagenomics, **BMC Bioinformatics**, **16**(1).
150. Liao, H., Shang, J., and Sun, Y., 2023. GDmicro: classifying host disease status with GCN and Deep adaptation network based on the human gut microbiome data, **Bioinformatics**.
151. Gou, W., Ling, C.w., and He, Y.v., 2020. Interpretable Machine Learning Framework Reveals Robust Gut Microbiome Features Associated With Type 2 Diabetes, **Diabetes Care**, **44**(2):358–366.
152. Larsen, P.E. and Dai, Y., 2015. Metabolome of human gut microbiome is predictive of host dysbiosis, **GigaScience**, **4**(1).

153. Sivakumar, A., K, S., and Jothi, J.A.A., 2023. Microbial Biomarkers Identification for Human Gut Disease Prediction using Microbial Interaction Network Embedded Deep Learning, **International Journal of Advanced Computer Science and Applications**, **14**(6).
154. , 2024, European Nucleotide Archive (ENA).
155. Franzosa, E.A., Sirota-Madi, A., and Avila-Pacheco, v., 2018. Gut microbiome structure and metabolic activity in inflammatory bowel disease, **Nature Microbiology**, **4**(2):293–305.
156. Schirmer, M., Franzosa, E.A., and Lloyd-Price, J.v., 2018. Dynamics of metatranscription in the inflammatory bowel disease gut microbiome, **Nature Microbiology**, **3**(3):337–346.
157. Nielsen, H.B., Almeida, M., and Juncker, A.S.v., 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes, **Nature Biotechnology**, **32**(8):822–828.
158. Imhann, F., Van der Velde, K.J., and Barbieri, R.v., 2019. The 1000IBD project: multi-omics data of 1000 inflammatory bowel disease patients; data release 1, **BMC Gastroenterology**, **19**(1).
159. Feng, Q. and Liang, S.v., 2015. Gut microbiome development along the colorectal adenoma–carcinoma sequence, **Nature Communications**, **6**(1).
160. Wirbel, J., Pyl, P.T., and Kartal, E.v., 2019. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer, **Nature Medicine**, **25**(4):679–689.
161. Yu, J., Feng, Q., Wong, S.H., and Zhang, D.v., 2015. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer, **Gut**, **66**(1):70–78.
162. Yachida, S., Mizutani, S., and Shiroma, H.v., 2019. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer, **Nature Medicine**, **25**(6):968–976.

163. Beghini, F., McIver, L.J., and Blanco-Míguez, A.v., 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3, **eLife**, **10**.
164. Shtossel, O. and Isakov, H.v., 2023. Ordering taxa in image convolution networks improves microbiome-based machine learning accuracy, **Gut Microbes**, **15**(1).
165. Chang, D., Gupta, V.K., Hur, B., and Cobo-López, S.a., 2024. Gut Microbiome Wellness Index 2 enhances health status prediction from gut microbiome taxonomic profiles, **Nature Communications**, **15**(1).
166. Jasner, Y.Y. and Belogolovski, A.v., 2020. MIPMLP – Microbiome Preprocessing Machine Learning Pipeline, **Frontiers in Immunology**.
167. Santoru, M.L. and Piras, C.v., 2017. Cross sectional evaluation of the gut-microbiome metabolome axis in an Italian cohort of IBD patients, **Scientific Reports**, **7**(1).
168. Nishida, A. and Inoue, R.v., 2017. Gut microbiota in the pathogenesis of inflammatory bowel disease, **Clinical Journal of Gastroenterology**, **11**(1):1–10.
169. Nishida, A.H. and Ochman, H., 2019. A great-ape view of the gut microbiome, **Nature Reviews Genetics**, **20**(4):195–206.
170. Barberio, B., Facchin, S., Patuzzi, I., and Ford, A.C.v., 2022. A specific microbiota signature is associated to various degrees of ulcerative colitis as assessed by a machine learning approach, **Gut Microbes**, **14**(1).
171. Alam, M.T. and Amos, G.C.A.v., 2020. Microbial imbalance in inflammatory bowel disease patients at different taxonomic levels, **Gut Pathogens**, **12**(1).
172. Plichta, D.R., Graham, D.B., Subramanian, S., and Xavier, R.J., 2019. Therapeutic Opportunities in Inflammatory Bowel Disease: Mechanistic Dissection of Host-Microbiome Relationships, **Cell**, **178**(5):1041–1056.

173. Franzin, M., Stefančič, K., Lucafò, M., Decorti, G., and Stocco, G., 2021. Microbiota and Drug Response in Inflammatory Bowel Disease, **Pathogens**, **10**(2):211.
174. Nomura, K., Ishikawa, D., and Okahara, K.v., 2021. Bacteroidetes Species Are Correlated with Disease Activity in Ulcerative Colitis, **Journal of Clinical Medicine**, **10**(8):1749.
175. Zhou, Y. and Zhi, F., 2016. Lower Level of Bacteroides in the Gut Microbiota Is Associated with Inflammatory Bowel Disease: A Meta-Analysis, **BioMed Research International**, **2016**:1–9.
176. Mah, C., Jayawardana, T., and Leong, G.v., 2023. Assessing the Relationship between the Gut Microbiota and Inflammatory Bowel Disease Therapeutics: A Systematic Review, **Pathogens**, **12**(2):262.
177. Kim, C.H., Lee, Y.U., and Kim, K.H.v., 2022. Comparison of Metabolites and Gut Microbes between Patients with Ulcerative Colitis and Healthy Individuals for an Integrative Medicine Approach to Ulcerative Colitis—A Pilot Observational Clinical Study (STROBE Compliant), **Diagnostics**, **12**(8):1969.
178. Wiredu Ocansey, D.K. and Hang, S.v., 2023. The diagnostic and prognostic potential of gut bacteria in inflammatory bowel disease, **Gut Microbes**, **15**(1).
179. Angriman, I., 2014. Relationship between pouch microbiota and pouchitis following restorative proctocolectomy for ulcerative colitis, **World Journal of Gastroenterology**, **20**(29):9665.
180. Markelova, M. and Senina, A.v., 2023. Association between Taxonomic Composition of Gut Microbiota and Host Single Nucleotide Polymorphisms in Crohn's Disease Patients from Russia, **International Journal of Molecular Sciences**, **24**(9):7998.
181. Hu, S., Png, E., and Gowans, M.v., 2021. Ectopic gut colonization: a metagenomic study of the oral and gut microbiome in Crohn's disease, **Gut Pathogens**, **13**(1).

182. Han, Y., Wang, B., and Gao, H.v., 2022. Insight into the Relationship between Oral Microbiota and the Inflammatory Bowel Disease, **Microorganisms**, **10**(9):1868.
183. Santana, P.T. and Rosas, S.L.B.v., 2022. Dysbiosis in Inflammatory Bowel Disease: Pathogenic Role and Potential Therapeutic Targets, **International Journal of Molecular Sciences**, **23**(7):3464.
184. Cisek, A.A., Szymańska, E., and Wierzbicka-Rucińska, A.v., 2024. Methanogenic Archaea in the Pediatric Inflammatory Bowel Disease in Relation to Disease Type and Activity, **International Journal of Molecular Sciences**, **25**(1):673.
185. Rinninella, E., Raoul, P., and Cintoni, M.v., 2019. What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases, **Microorganisms**, **7**(1):14.
186. Ahn, J., Sinha, R., and Pei, Z.v., 2013. Human Gut Microbiome and Risk for Colorectal Cancer, **JNCI: Journal of the National Cancer Institute**, **105**(24):1907–1911.
187. Montalban-Arques, A. and Scharl, M., 2019. Intestinal microbiota and colorectal carcinoma: Implications for pathogenesis, diagnosis, and therapy, **EBioMedicine**, **48**:648–655.
188. Sherafat, S.J., Azimirad, M., and vd., M.A., 2019. The rate and importance of Clostridium difficile in colorectal cancer patients, **Gastroenterology and Hepatology from Bed to Bench**, **12**:358–363.
189. Elahi, Z. and Shariati, A.v., 2023. Association of Lactobacillus, Firmicutes, Bifidobacterium, Clostridium, and Enterococcus with colorectal cancer in Iranian patients, **Heliyon**, **9**(12):e22602.
190. Keku, T.O., Dulal, S., Deveau, A., Jovov, B., and Han, X., 2015. The gastrointestinal microbiota and colorectal cancer, **American Journal of Physiology-Gastrointestinal and Liver Physiology**, **308**(5):G351–G363.

191. Hong, Y., Song, D., Liu, N., and Wang, H., 2023. Development and treatment of colorectal cancer: Insights from multi-kingdom microbiota, **Aging and Cancer**, **4**(1):21–40.
192. Liu, J., Dong, W., and Zhao, J.v., 2022. Gut microbiota profiling varied during colorectal cancer development in mouse, **BMC Genomics**, **23**(S4).
193. Xu, Y., Zhao, J., and Ma, Y.v., 2023. The microbiome types of colorectal tissue are potentially associated with the prognosis of patients with colorectal cancer, **Frontiers in Microbiology**, **14**.
194. Hexun, Z., Miyake, T., and Maekawa, T.v., 2022. High abundance of Lachnospiraceae in the human gut microbiome is related to high immunoscores in advanced colorectal cancer, **Cancer Immunology, Immunotherapy**, **72**(2):315–326.
195. Zhao, L., Cho, W.C., and Nicolls, M.R., 2021. Colorectal Cancer-Associated Microbiome Patterns and Signatures, **Frontiers in Genetics**, **12**.
196. Taglialegna, A., 2023. Commensal bacteria fight colorectal cancer, **Nature Reviews Microbiology**, **21**(5):276–276.
197. Thomas, A.M. and Manghi, P.v., 2019. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation, **Nature Medicine**, **25**(4):667–678.
198. Ma, B., Wang, Y., and Ye, S.v., 2020. Earth microbial co-occurrence network reveals interconnection pattern across microbiomes, **Microbiome**, **8**(1).

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı, Soyadı : Umay Gülfem ELGÜN ÇİFTÇİOĞLU
Uyruğu : Türkiye (T.C.)
E-posta :

EĞİTİM

Derece	Kurum	Mez. Yılı
Lise	Nuh Mehmet Küçükçalık Anadolu Lisesi, Kayseri	2011
Lisans	Erciyes Üniversitesi, Bilgisayar Mühendisliği	2017
Yüksek Lisans	Erciyes Üniversitesi, Bilgisayar Mühendisliği	2019

İŞ DENEYİMLERİ

Yıl	Kurum	Görev
2022-2022	Hasan Kalyoncu Üniversitesi, Yazılım Mühendisliği Bölümü, Gaziantep	Araştırma Görevlisi

YAYINLAR

SCI, SSCI, AHCI İndekslerine Giren Dergilerde Yayımlanan Makaleler

- Ciftcioglu, U. G. E. & Nalbanoglu O. U., 2024. DeepGum: Deep feature transfer for gut microbiome analysis using bottleneck models, Biomedical Signal Processing and Control, 1746-8094, 91, 8.
- Ciftcioglu, U. G. E. & Nalbanoglu O. U., 2025. Disrobiom: A novel approach

to discover robust biomarkers from gut microbiome datasets with deep-learning algorithms, Biomedical Signal Processing and Control, 1746-8094, 100, 106935.

SCI, SSCI, AHCI İndekslerine Giren Dergilerde İncelemedeki Makaleler

- Ciftcioglu, U. G. E. & Nalbanoglu O. U., Deeptaxim: Comprehensive Classification Analysis for Taxonomic Datasets Using Image-Based Deep-Learning Models, IEEE Transactions on Computational Biology And Bioinformatics. (Hakem incelemesinde)

Hakemli Kongre/Sempozyumların Bildiri Kitaplarında Yer Alan Yayınlar

- Ciftcioglu, U. G. E. & Nalbanoglu O. U., Mikrobiyom Verilerinde Öznitelik Seçimi Yöntemlerinin Karşılaştırmalı Analizleri, V. Bilisel Uluslararası Gordion Bilimsel Araştırmalar Kongresi, Ankara, Türkiye, 8-9 Aralık, 2024.
- Ciftcioglu, U. G. E. & Nalbanoglu O. U., Derin Öğrenme Modelleri ile İnvaziv Olmayan Fenotip Tahmini, V. Bilisel Uluslararası Gordion Bilimsel Araştırmalar Kongresi, Ankara, Türkiye, 8-9 Aralık, 2024.